ESTIMATION OF THE NUMBER OF MICROBIAL SPECIES COMPRISING A POPULATION

THESIS

Melanie R. Slattery, Captain, USAF

AFIT/GAM/ENC/08-03

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force Department of Defense, or the United States Government.

AFIT/GAM/ENC/08-03

ESTIMATION OF THE NUMBER OF MICROBIAL SPECIES COMPRISING A POPULATION

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Applied Mathematics

Melanie R. Slattery, BA

Captain, USAF

March 2008

AFIT/GAM/ENC/08-03

ESTIMATION OF THE NUMBER OF MICROBIAL SPECIES COMPRISING A POPULATION

Melanie R. Slattery, BA
Captain, USAF

Approved:

/signed/

Samuel A. Wright, Maj., USAF (Chairman) date

/signed/

Charles A. Bleckmann (Member) date

/signed/

Stephanie A. Smith (Member) date

Abstract

The purpose of this research was to evaluate the appropriateness of using non-parametric estimators, specifically the Chao1, ACES, and Jackknife methods, for estimation of the number of unique species comprising a population. It goes on to develop a parametric method for the above stated problem. This research consisted of creating diverse populations, with known numbers of species, and applying the aforementioned non-parametric and parametric methods to samples drawn from the constructed populations. The parametric fitting of several different distributions to the sample data, including the lognormal, gamma, and Weibull was considered. Both types of methodologies were then applied to sample data from constructed wetlands, where little is known about the overall population size and species composition (number of unique species in the population). This research attempted to identify the underlying population distribution of the wetlands (via fitting of parametric curves to the sample), as well as focused upon demonstrating that the use of parametric methods were more apt to provide better results in estimating the number of species in a natural population.

This research discovered the use of the non-parametric methods, developed originally for the use of smaller well-defined populations (Chao1) or computer debugging (ACES) was not appropriate for species estimation. The use of these methods resulted in lower bounds, which were several standard deviations away from the true number of species, for the contrived populations. This research found applying a parametric method was more accurate in representing the truth. A comparison of the two different approaches to species estimation and the advantages of using a parametric method over a non-parametric method are discussed as well.

Acknowledgments

I would like to thank my family. I would like to thank my faculty advisor, Maj Samuel Wright, for his guidance and patience throughout the thesis process. I would also like to thank Dr. Charles Bleckmann and Dr. Stephanie Smith for allowing me the opportunity to study the wetlands from a statistical perspective. In addition, I would like to thank all of the aforementioned for sharing their valuable knowledge and expertise with me; it was essential to my success.

Melanie R. Slattery

Table of Contents

List of Figures

Figure Page

Figure Page

List of Tables

Table Page

List of Equations

PARAMETRIC ESTIMATION OF NUMBER OF SPECIES COMPRISING A POPULATION

I. Introduction

Background

The current state of the environment continues to become an increasing area of focus and concern, as researchers look for ways to negate the detrimental impact humans have had on the environment. At the forefront of efforts to improve its state are the techniques which are currently used to remediate the environment. Although there are different methods available, many are too costly to be useful for widespread application. Bioremediation is a natural method which holds much promise in its ability to efficiently, effectively decontaminate polluted areas.

This particular method relies upon microbes found within the soil to decompose carcinogenic and pollutant materials. Consequently, the composition of the microbial population in soil is receiving a great deal of attention. The theory behind this is that if the species of microbes known to be effective as remediation agents can be determined, then perhaps these specific species can be introduced into heavily polluted areas and successfully begin and/or complete a remediation process. In order to know what species play an important role in remediation, scientists first must learn the number of species that comprise a population. This biological categorization may lead to a functional categorization, which will allow biologists to determine the importance of specific species in certain roles. However, this seemingly simple task of determining the composition and number of species is, in fact, a momentous problem.

The composition of the underlying population of the microbes is, to date, unknown, as are the exact number of species which reside in the soil. Methods for the estimation of the number of species which are currently used include both parametric and non-parametric estimation techniques. There are advantages and disadvantages to the use of parametric versus non-parametric techniques. Some believe that non-parametric estimators will afford the most accurate estimations. Others believe that the parametric estimators will lend themselves better to estimation.

Problem Statement

The estimation of species diversity is a matter of interest in the bioremediation process. The goal of this research is to discern the most appropriate method of estimating the number of species found in a particular population. This research makes use of data collected from constructed wetlands as an example of practical application of the most appropriate method.

Research Objectives

There are a plethora of methods used to estimate the diversity of microbial populations in the soil. This thesis aims to derive a best estimator for the use of examining the aforementioned problem.

Before the process of derivation can begin, it is necessary to understand and review the parametric and non-parametric methods which are commonly used for population estimation. This thesis aims to review the advantages and disadvantages of

the most commonly used estimators. It then continues by seeking the derivation or modification of a best estimator.

Research Question

Currently, non-parametric estimators are touted as best for estimating the number of species in a microbial population. The overarching question, for the purpose of this study is: Are the non-parametric methods currently used for species estimation, in fact, appropriate for use on estimating the number of species comprising a natural microbial population, or does a better alternative exist?

This thesis begins first by answering the question of the correctness in applying the commonly used non-parametric methods in this situation. This thesis then continues on to suggest a more appropriate alternative for wetlands soil species estimation.

Thesis Organization

Chapter 2 begins by examining the breadth of research currently available regarding microbial population species estimation. There are varying viewpoints as to applying a non-parametric versus parametric method to the problem at hand. This chapter attempts to examine the differing opinions and offer the positive and negative aspects of both arguments.

Chapter 3 contains the methodology used in this thesis. The undeniable fact that so many unknowns surround this problem makes it difficult to assess the truth. This portion of the thesis delves into the creation of several populations, which when sampled produced (in some cases) samples similar to the real-world data collected from the

constructed wetlands. Also included were populations from which samples were comprised of mostly unique species. Particular emphasis was placed upon creating populations in which many of the sampled individuals were seen as unique.

Chapter 4 delves into the analysis and results produced by the methodology used in this thesis. This chapter will focus upon reviewing the results of using each method on several different populations. Comparison charts will be included to further demonstrate the results from the different methodologies.

Chapter 5 introduces the conclusions and recommendations of this thesis. These conclusions will be based upon both the application of a parametric method to the fabricated populations as well as to the real-world constructed wetlands data. Suggestions for future research are also included in this chapter.

II Literature Review

Chapter Overview

The purpose of this chapter is to review relevant research in the area of composition biodiversity estimation. According to the Stanford Encyclopedia of Philosophy, compositional biodiversity is the ability to separate organisms into categories based upon structural composition (as opposed to a classification based upon an organism's functionality). This categorization can itself be studied from different aspects. Biodiversity can be examined in terms of the number of different species present in the community, often referred to as a community's richness. Alternatively, biodiversity can also be viewed in terms of coverage. A measure of coverage is evenness, which describes how many of each species is present in the community. Biodiversity can be viewed as a combination of both richness and evenness, a method which, perhaps, provides the most information about the community.

There are a plethora of algorithms available for estimating the different aspects of biodiversity. Some estimators, such as the Chao1 method, strictly focus on richness data. Methods such as these will focus on the number of times each species is found in the total sampling effort. For example, the Chao1 estimator considers the number of species that are only seen one or two times; all other information is not accounted for in the calculation (Chao; 1987). Other estimators, such as Simpson's index, pull both richness and evenness information from the sample. The calculation of Simpson's index takes into account the number of individuals of each species in the sample as well as the number of species which appear in the sample (Offwell Woodland and Wildlife Trust; 2007).

Furthermore, these existing methods may also be partitioned depending on assumptions made regarding the underlying population. The estimators may be parametric, non-parametric, or a combination of the two. Each of these classes has advantages and disadvantages associated with their use in estimating diversity of microbial population.

Description

Many of the currently used diversity estimation methods were first developed to describe populations of macro-organisms (for example species of birds, foxes, etc.). Populations of macro-organisms, though they may be comprised of hundreds of different species, are generally orders of magnitude smaller than the number of species associated with microbial populations.

This difference in magnitude is evidenced by examining accumulation curves (of various types of organisms) which are used to visually express the relationship between the sampling effort and the number of species found throughout that effort (Hughes et al.; 2001). Populations which are comprised of many of the same species (such as birds or foxes) will reach an asymptote quickly, whereas populations comprised mostly of singletons (the only member of the species) will appear to be more linear (Hughes et al.; 2001). Figure 1 provides an example of two accumulation curves. The line represented by the filled-in circles is an example of an asymptotic accumulation curve; the line represented by the open circles is an example of a non-asymptotic accumulation curve (Dove and Cribb; 2006:569).

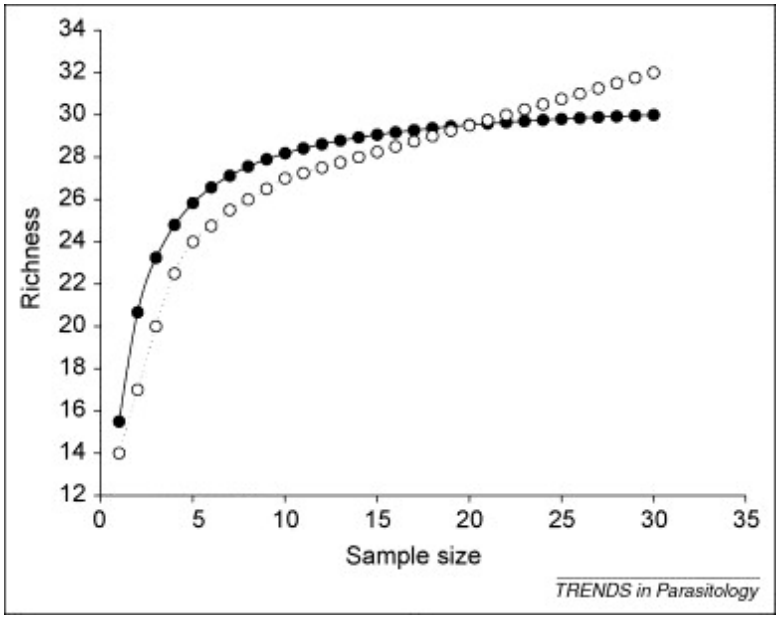

Figure 1. Examples of Asymptotic and Non-Asymptotic Accumulation Curves (Dove and Cribb; 2006:569)

The theory behind the use of accumulation curves is, given that a community has a finite number of species, and, given that the population is able to be sampled abundantly (a key assumption), eventually all the species will be seen by the researchers and the curve will reach an asymptote. Consequently, populations consisting of fewer species will require less time and effort in terms of sampling (smaller sample sizes will predict the diversity of the sampled community). However, a population comprised of many different species, such as some insects or microbes, will require a greater sampling effort, and the accumulation curve will be appear to be linear if the sample is not large enough to infer the number of species in the population.

Although accumulation curves may be practical for macro-organisms, since microbial populations (in many instances) are too diverse to count in this way,

accumulation curves are not ideal and will not give much information regarding the population, as stated above, the curve will appear to be linear (Hughes et al.; 2001). It is therefore necessary to use a diversity estimator to make inferences about the population composition.

The same algorithms which have been developed for (more or less) well-defined populations have been directly applied to microbial populations where there is a great amount of uncertainty surrounding the number of species in a particular population. One of the fundamental problems with this type of direct application is that, for at least a few of the diversity indices (for example, the Shannon and Simpson indices) the total number of species needs to be known (Chao and Shen; 2003; Collins et al.; 1995). To date, however, the number of microbial species residing in the soil is unknown (or there would be no need for diversity estimators). This issue is further complicated by the length of time required to sample and identify microbial species. For example, when using certain diversity measures, such as the Chao1 estimator, the estimate will be correlated with the sample size until the square root of two times the entire richness is reached ($\sqrt{2*richness}$) (Hughes et al.; 2001: 4401). With current estimates of bacteria soil population ranging from 3,000,000 to 5,000,000,000 per gram of soil the effort involved in achieving a sample which will accurately reflect the population quickly grows (American Society of Agronomy; 1977). Realistically, with the limited resources available, it becomes nearly impossible, for an extremely large and diverse population `to collect a sample size that will result in an uncorrelated sample.

Therefore, an estimator which can accurately predict the number of microbial species with the available sample (which in most cases may not be a sample of sufficient

size to produce an uncorrelated estimate using non-parametric methods) is crucial to answering many questions regarding microbial soil diversity.

Relevant Research

Before comparing or discussing the different methods currently used for population estimation, it is first necessary to discuss the more commonly used estimators. The more commonly used methods include the use of extrapolation from an accumulation curve as well as the use of non-parametric and parametric estimation. The non-parametric estimators which will be discussed are the Chao1 method, the abundance coverage estimator (ACES), and the Jackknife method. The parametric estimators which will be discussed will be based upon the population distributions which are thought to be those which most accurately describe biological populations. These include the lognormal, broken stick, the geometric, and, most recently, the Pareto (Hong et al.; 2006: 118).

Accumulation curves may themselves be used to estimate the number of unseen species composing a particular population; however, it is important to note the way in which samples are added to the curve (Colwell and Coddington; 1994). The difficulty of the method of introducing samples to the curve may be quickly overcome by ensuring randomization of the sample order (Colwell and Coddington; 1994). The curve may be computed either by rarefaction or a method described as a random placement curve (Colwell and Coddington; 1994). Once the curve is created, based upon the assumption that the curve represents a uniformly sampled process, extrapolation may be completed by using either an asymptotic or a non-asymptotic approach (Colwell and Coddington;

1994). Many different models to extrapolate from the curve exist which create an inherent problem with the use of this method for estimation. This is because the application of different extrapolation methods may result in different answers for any given sample (Colwell and Coddington; 1994). Additionally, different models may be more appropriate for different populations. Although one model may be used to estimate a particular population, it may be inappropriate to apply the same model to a dissimilar population (Colwell and Coddington; 1994).

One of the most commonly used non-parametric methods is one developed by Anne Chao in 1987. It is commonly referred to as the Chao1 method and is based upon the use of capture-recapture sampling, which is often used in studying macro-organisms. The capture-recapture method is a common method for studying the size of populations comprised of macro-organisms, such as birds or foxes. Individuals from the population are captured, tagged, and released. This process is repeated for a set number of days. Any tagged individual who is trapped again is considered to be recaptured (Chao; 1987).

The Chao1 method is designed for situations in which there are many individuals which are seen only a small number of times (one or two) (Chao; 1987). To use in populations of microbes, a microbe is recovered from a sample and is identified (by species name or otherwise identified if it is a previously unknown species) and placed in a "bin." This is equivalent to capturing a species. Each subsequent microbe is identified. Microbes that are identified as belonging to the same species (for the purpose of this paper microbes with 97 percent genetic similarity of the 16s ribosomal ribonucleic acid (16s RNA), which is a piece of an organism's RNA which uniquely defines it as a species, are considered to belong to the same species) are placed in the same bin. This is

considered to be equivalent to re-capturing. If they do not match a new bin is created and a new species is captured (Chao; 1987).

The theory behind the process follows. Each bin is assigned a probability from some probability distribution *F* which allows for unequal probability of capture (as opposed to the uniform distribution which would require equal probability of capture for each bin). There will be a certain number of distinct species captured in the experiment (denoted by *S*). Additionally, f_k represents the number of microbes captured *k* times in the samples. It then follows that the entire number of unique species (*N*) will be equal to the number of distinct species found plus some number of species which are never observed in the sample. It has been proven that the joint unconditional distribution of $(f_0, f_1, ..., f_t)$ is multinomial (Burnham and Overton; 1978:628). After the creation of a cumulative distribution function and some manipulation and substitution of variables the following equation for an estimator of *N*, the total number of species in a population:

$$\hat{N} = S + f_1^2 / (2 f_2) \qquad (1)$$

where *S* is the number of distinct species caught in the sample, f_1 is the number of singleton species caught in the sample, and f_2 is the number of species with exactly two members captured in the sample (Chao; 1989).

The only information included in the estimator from the sample are the number of species which are identified either one or two times. Species appearing more often than this are considered negligible. However, if the information regarding the more abundant species is in fact non-negligible then this estimator is, at best, a lower bound of the

number of species found in the population (Chao; 1987). This point seems to be often overlooked in studies and consequently it is often stated that the estimator often significantly underestimates the diversity of the community (Kemp and Aller; 2004). Another overlooked essential point is if the capture probabilities stem from a distribution which results in a relatively large average capture probability this method is not acceptable (Chao; 1987).

The abundance-based coverage estimator (ACES) was derived by Anne Chao and Mark Yang. It is another non-parametric estimator frequently applied to the problem of microbial biodiversity. This method was originally intended for estimating software errors, for which a Poisson process is used to describe the rate of the occurrences of the *i*th error (Chao and Yang; 1993: 193). Unlike the Chao1 method, this particular estimator takes into account species which are identified as rare as opposed to considering only the singleton and doubleton species. Rare is defined as identifying the species 10 or fewer times within the sample (Chao and Yang; 1993).

The sample coverage is a vital piece of information when using this estimator. Sample coverage addresses random samples from multinomial populations and is defined by Chao and Lee as "the sum of the probabilities of an observed class," (Chao and Lee; 1992: 210). Other essential quantities needed for the calculation are the mean of the Poisson rates for all errors (species capture rate):

$$\bar{\lambda} = \frac{\sum \lambda_i}{N} \qquad (2)$$

and the coefficient of variation, denoted as γ, where:

$$\gamma = \frac{\sqrt{\sum \left(\lambda_i - \overline{\lambda}\right)^2 / N}}{\overline{\lambda}} \tag{3}$$

(Chao and Yang; 1993: 194). Additionally, the number of unknown errors (species) which have not been detected can be denoted as $f_0 = N - D$, where D is defined as the number of distinct errors (species) which have been detected in the sample. After combining and manipulating the equation the resultant estimator is:

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{f_1}{\hat{C}} \hat{\gamma}^2 \tag{4}$$

where $\hat{C}$, the estimator for sample coverage is:

$$\hat{C} = 1 - f_1 / \sum i f_i \tag{5}$$

and f_1 is the number of errors (species) occurring only once (Chao and Yang; 1993).

The Jackknife estimator is the final non-parametric method examined in this research. The generalized Jackknife method was first applied to the problem of species estimation by Burnham and Overton. Like the Chao1 and ACES methods, this method does not require an assumption of equal capture probabilities among species (Burnham and Overton; 1978:625). This method is essential to the derivation of the Chao1 method in that the Chao1 method is a variation of the derivation of the jackknife estimator

(Chao; 1987:784). This method takes the joint, unconditional distribution function of the species in a population and applies the generalized jackknife technique (Burnham and Overton; 1978:629). The resulting generalized equation, whose *k*-th order form is given by:

$$N_{Jk} = \sum_{i=1}^{t} a_{ik} f_i \qquad (6)$$

and must be derived for each N_{Jk} as no simplified result of the equation exists (Burnham and Overton;1978:628). According to research, this estimator reportedly underestimates the number of species if the population is comprised of many species which are identified only a few times (Chao; 1987:784). The first and second order estimators are included in this research as a comparison with the Chao1 and ACES methods to assess how well, with the data presented in this thesis, each is able to estimate the number of species in an extremely diverse population.

Another class of estimators exist which are also often used to describe microbial population composition. These are the parametric estimators. In order to use a parametric estimator, assumptions must be made about the underlying population distribution. In the case of estimating the number of microbial species present in the soil, an assumption must be made regarding the distribution of all the species thought to be present in the soil. In many cases regarding biological/ecological data, this is done by assuming the number of individuals versus the number of species takes on a continuous lognormal or Poisson lognormal distribution (Hughes et al; 2001:4401). Once the

assumption regarding the total population has been made, the sample data is then fit to the same distribution. This then allows for estimating the parameters of the curve and the total number of species can therefore be estimated (Hughes et al; 2001:4401).

However, some of the assumptions made regarding use of the continuous lognormal model are troublesome. For example, different groupings of abundance categories (since discrete data is used, rounding is necessary and the method used for rounding may result in the creation of different categories within the model) will yield different estimates (log base *j* will differ from log base *k* assuming *j* is not equal to *k*) (Colwell and Coddington; 1994:108). Additionally, the placement of singletons within the groupings gives cause for concern and may ultimately bias the estimate of the mean (Colwell and Coddington; 1994:108).

The Poisson distribution may fair better as it is a discrete distribution meant for data such as species identification which is discrete in form. This model however, does not appear to be used as frequently as expected. Colwell and Coddington suggest this may be due to the difficulty associated with fitting the model (Colwell and Coddington; 1994:108).

Some of the more common distributions thought to apply in this situation, in addition to the logarithmic and Poisson distributions are the geometric and broken stick models. The geometric can be used to describe communities dominated by a few species with others being rare (this designation is based upon resource allocation, where resources are considered to be those items required for microbial survival) (Kemp and Aller; 2004:116). The broken stick can be described as a community in which resources are allocated evenly throughout resulting in a fairly even community composition (Kemp

and Aller; 2004:116). In more recent literature, the use of the gamma and inverse Gaussian have also been explored (Hong; 2006).

Recently, another distribution has been regarded as having potential in this area. The Pareto distribution may be used to explain extreme populations. In other words, it may be used to describe populations which are composed mostly of rare species or mostly of abundant species (Hong; 2006:118). Although this model shows promise in its ability to estimate the number of species present in a population, it has not yet seen great use. It has been suggested that this promising distribution be studied further in its applicability to this problem.

A study comparing six different parametric methods found that the accuracy of the estimate depends upon how the species are identified (Hong; 2006:118). For example, if organisms had to be only 80 percent genetically similar in order to be classified as the same species, the Inverse Gaussian distribution best described the population, whereas, if the organisms required 99 percent similarity, the lognormal model best described the data (Hong; 2006:119). The conclusion reached in this study was that parametric methods may in fact be a more appropriate way to estimate the total number of species in a population.

Summary

Though there are a multitude of ways to estimate the number of species comprising a particular population, they can be grouped into three general classes:

extrapolation from an accumulation curve, use of non-parametric estimators, and use of parametric estimators. To date, the non-parametric methods have dominated the field, but recent research suggests that this may be inappropriate (Hong; 2006).

Non-parametric methods have the advantage of not making any assumptions regarding the underlying population which is being sampled. A major drawback with the use of these algorithms is they appear to be correlated with the sample size and consequently appear to underestimate the total number of species which are present (Hong; 2006:119, Colwell and Coddington; 1994:111).

The Chao1 estimator, one of the most frequently used methods, is intended to be a lower bound (Chao; 1994). Additionally, if the estimator is used in its original form and there are no doubletons present the estimator is unusable (Chao; 2006).

Non-parametric estimators as a group tend to give varying answers when applied to the same problem, with some appearing more correlated with sample size than others. Several studies applied multiple non-parametric methods which computed differing numbers for total species (Colwell; 1994, Kemp; 2004, Kemp and Aller; 2004, Hughes; 2001, Chao; 1984).

Parametric estimators are less likely to be used when estimating the number of species. This may be because certain assumptions must be made regarding the underlying population distribution. Nonetheless, a number of different distributions have been applied to this problem. Although the more commonly used are the lognormal and the Poisson, it appears that other less known distributions may be better suited to explain microbial population distribution, such as the Pareto (Hong; 2006).

According to the biological literature, the sample size has a significant impact upon which estimator should be applied. It is important to understand that the microbial communities seem to be comprised of enormous numbers of species. As such, it is neither time efficient nor practical to sample exhaustively. Thus the need for an estimator which is as accurate and unbiased as possible is essential.

III. Methodology

Chapter Overview

This chapter will discuss the methodology used to evaluate the commonly used non-parametric estimators as well as discuss the manner in which parametric methods were applied to the problem. As previously discussed, the underlying question is the determination of the best method to use for estimating the number of species in a population (specifically in this case, a population of enormous size). The first step was to create populations where the truth, regarding the number of species in the population, is known. The next step was to evaluate the most commonly used non-parametric methods on the contrived data. This was followed by the application of parametric curve fitting to the samples. The latter two steps were applied to data collected from the constructed wetlands. Figure 2 gives a pictorial representation of the aforementioned methodology.

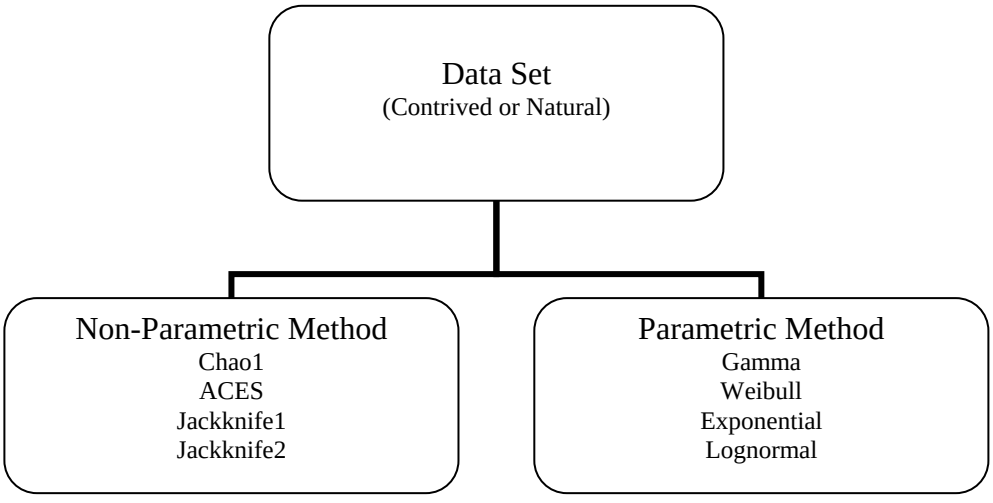

Figure 2. Methodology

Data Collection (Contrived Population)

One of the most difficult aspects of this problem was the fact that the number of species residing in the soil is unknown and is almost certainly a number of large magnitude. It was therefore difficult to assess how well common estimators will perform if there is so little known about the truth.

It was considered appropriate, then, to evaluate the various estimators through their use on well-defined populations. That is to say, populations in which the total number of species were known. Although this has been done before (Chao uses comparison tables in several of her papers to evaluate her estimators), the populations used for the evaluations are, quite likely, several orders of magnitude smaller than that of a constructed wetlands. Chao used examples of taxicab driver and cottontail rabbit populations to assess the accuracy of her method (Chao; 1987:787). These populations numbered 420 and 135 individuals respectively. Soil-based microbial populations are generally thought to be orders of magnitude larger than the populations used by Chao. Consequently, it was imperative to create larger populations in order to evaluate the performance of different estimators.

Population Creation

Populations were created by programming a generic gamma probability distribution function into Matlab in conjunction with Matlab's built-in Probability Distribution Toolbox.

The equation used for the gamma distribution was:

$$f(x) = x^{\alpha-1}\frac{\beta^{\alpha}e^{-\beta x}}{\Gamma(\alpha)} \tag{7}$$

where the equation used for incomplete gamma function was:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt \tag{8}$$

In order to create a population of the desired magnitude, parameters were chosen so that with large numbers of "species" (on the order of 10,000) the density of the function, when multiplied by 10^{10} results in at least 0.5. This was so that when rounded, the number of species considered contains at least one individual. After determining the alpha and beta parameters for use with the gamma distribution, the parameters were then run through a program which created bins, with each bin containing at least one individual. Individuals which were in the same species were placed in a bin. The number of bins corresponded to the number of species in the population. The total population size was determined by summing up the number of individuals in each bin across all the bins. The following was the equation used for the summation:

$$Pop = \sum_{i=0}^{N} bin(i) \tag{9}$$

A crucial step in the creation of a population was the ability to determine which individuals belong to which species. That is to say, the previous steps gave a method for determining the number of individuals in each category, but no way to determine which species a particular sampled individual belongs to. This problem was solved by assigning numbers to each individual (e.g., numbers 0 through 1,000 would belong to species A, 1,001 through 1,500 would belong to species B). The application of a cumulative summation function to the bins assigned each individual to a bin and completed the process of creating a population.

The programming of each step in Matlab enabled the procedure of population generation to be repeated any number of times, allowing the population parameters and size to be easily changed. This allowed for the manufacture of several different populations.

Sampling of Contrived Populations

Random samples were drawn from the population using Excel. Excel's randbetween function running from one to X (where X was the total size of the population) was used to sample individuals. The randbetween function allowed for random integers, as opposed to decimal values, to be drawn. It is important to note that this function is based upon sampling with replacement. This introduces a problem in that there is a possibility that the same number can be drawn twice, an issue unique to the manufactured populations and one not found in natural, biological sampling of microbes from the soil. Since the drawing of the same number was not acceptable for the purpose of this thesis, this problem was overcome by drawing a larger sample than required and

taking the first *n* unique numbers (*n* being the desired sample size). Each sample was then imported into Matlab for categorization.

The Matlab code used for the procedure is included in an Appendix A to the document. The total numbers of sampled individuals drawn from each species was counted. The sample was then considered ready for analysis.

Wetlands Data Collection

Data from the constructed wetlands was collected (Leon; 2008) and provided for use in this thesis. The data was supplied in a form such that the methods in question (both the commonly used non-parametric and the proposed parametric methods) were able to be readily applied. Figure 3 depicts a break-out of the wetlands data at a group level (species have not been yet identified).

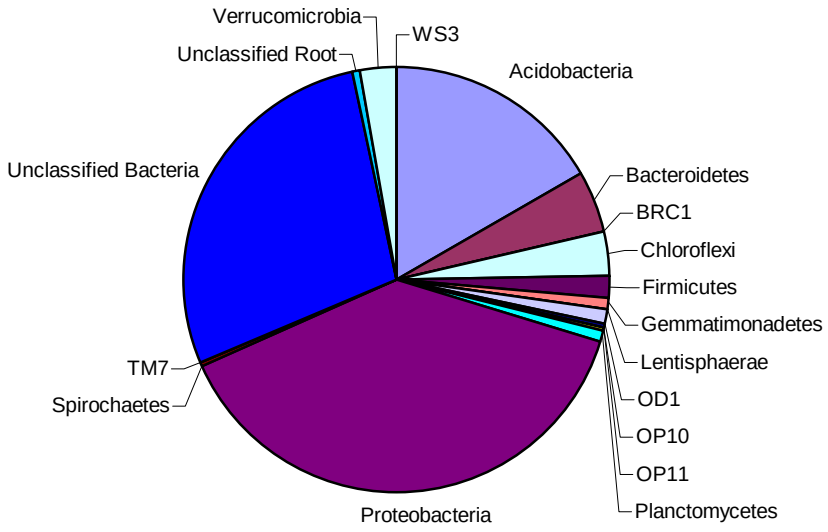

Figure 3. Wetlands Soil Data (Leon; 2008)

Non-Parametric Method Application

The following non-parametric methods were considered in this thesis: Chao1, ACES, Jackknife 1, and Jackknife 2. Of note is that the two most commonly used methods are Chao1 and ACES. The formulations for these four methods were programmed into Matlab for use in this thesis. Chapter 2 contains the equations used for all the non-parametric methods considered in this thesis. Additionally, the code included the equation for the variance of the Chao1 method. The Matlab code used is included in Appendix B to this document. The code was verified with the examples provided by Chao (Chao, 1987). Furthermore, code verification, this time on the wetlands data, was performed with a widely-used computer software program (DOTUR).

Parametric Method Application

Parametric curve fitting was applied to all samples with the use of EasyFit software. This allowed for evaluation of multiple distributions (both continuous and discrete) compared with the sample data. The commercial software ranked each distribution according to the Kolmogorov Smirnov goodness of fit test. The purpose of the test was to determine if the data in question comes from a specific probability distribution (NIST; 2008). The null hypothesis states that the data does indeed come from a specific probability distribution; whereas the alternative hypothesis is that the data does not come from the specific probability distribution (NIST;2008).

The test statistic for the Kolmogorov Smirnov test is defined as:

$$D = \max_{1 \le i \le N} \left(F\left(Y_i\right) - \frac{i-1}{N}, \frac{i}{N} - F\left(Y_i\right) \right) \tag{10}$$

The null hypothesis is then rejected or failed to be rejected at several alpha levels (all calculations were performed automatically by the EasyFit software).

In actuality, the data being studied in this problem follows a discrete, not a continuous distribution. To overcome this problem, the top-ranked distributions were discretized. There are a multitude of methods commonly accepted and used to approximate discrete distributions from continuous ones. The method chosen for this particular thesis was applying Riemann sums to the distribution as programmed into Matlab.

The code used for the discretizing is included in Appendix C to this thesis. Extrapolation was the next step. The data was multiplied by 10^{10} in the case of the contrived populations. For the wetlands soil data, the size of the wetlands was multiplied by the number of individuals thought to exist in a gram of soil to get a lower estimate for the number of bacteria residing in the entire wetlands.

Limitations

The first limitation, which has been repeatedly emphasized throughout this thesis, is that there is not currently a solid estimate as to how many microbes are living in the soil. This limitation makes the assessment of estimators on actual data difficult. Contrived populations will give the advantage of knowing the truth (and thus a way to evaluate estimators), but may not reflect the shape and size of the actual data in question.

Another limitation is the choice of a parametric estimator, where assumptions must be made about the underlying population. Again, since so little is known about the

truth of microbial populations, making assumptions about the bacterial composition is an obvious concern.

A further concern regarding this limitation revolves around the fact that non-parametric methods although invariant to population size are variant to sample size. This means that the size of the underlying population, when using a non-parametric method, will not change the estimate. However, the size of the sample will have an impact on the resultant estimate. In the case of using parametric methods, the reverse is true. The sample size should not have an effect on the use of the method, but the overall population size will have an effect (when extrapolated). This implies that parametric methods are variant to population size but are invariant to sample size. The non-parametric portion of this limitation will be addressed in Chapter 4, illustrated by an example using both a contrived population and the wetlands data.

Summary

This chapter explains the methodology used to begin answering the question regarding species estimation. It delves into the creation of large populations and the reasoning behind their use. Both non-parametric and parametric methodologies are described with regard to the contrived data as well as their application to the real-world data. Limitations of the overall problem as well as those of the proposed method are discussed as well.

IV. Results and Analysis

Chapter Overview

This chapter introduces the results and analysis produced by this research. First and foremost it addresses the question of the appropriateness of the use of the more commonly used non-parametric methods in practical application of microbial species population estimation. This is accomplished by applying the methods to the fabricated populations described in the previous chapter. This chapter then analyzes the results of using a parametric approach to both the fabricated data as well as the real-world wetlands data. Each contrived population will be discussed in detail, as will the results of the analysis of the wetlands data.

Results of Research

Six different sample populations were created and analyzed using both the non-parametric as well as the parametric methods. The size and number of species varied from population to population. The underlying distributions for the contrived populations were all two-parameter gamma distributions. The truth regarding the underlying population distribution of species for the wetlands is unknown.

Population 1

Population 1 was created by using a gamma distribution with an alpha parameter value of 1 and a beta parameter value of 200. Population densities were multiplied by 10^{10} to ensure a large population size. The resulting population contained 997,501,983

individuals belonging to 3,223 different species; the resultant population is smaller than one billion due to rounding and discretizing continuous data. A random sample of size 1,000 was taken and categorized as described in Chapter 3 of this document. Table 3 contains the pertinent population characteristics.

Table 1. Population 1 Specifications

Probability Distribution	Alpha Parameter	Beta Parameter	Population Size	Number of Species	Sample Size
Gamma	1	200	997,501,983	3,223	1,000

The sample resulted in 453 unique species, of which 227 were singleton species. Figure 4 depicts a graphic representation of the sample. Due to the narrow width of each category and the number of unique species, the columns appear to be connected, when, in fact they are not.

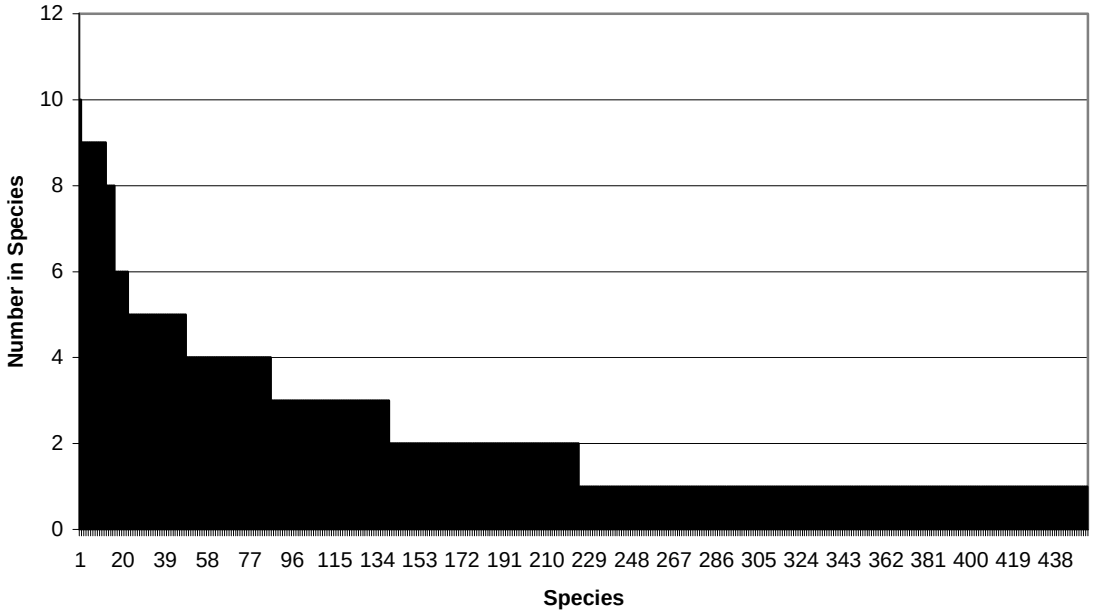

Figure 4. Sample Data Population 1

This sample is similar to the wetlands soil data set in that a majority of the sample is comprised of singleton and doubleton species. Unlike the wetlands soil data, this particular sample resulted in a sample comprised of rare species (no sampled species contained more than 10 individuals).

The four non-parametric methods were applied to the population data. Table 2 contains the results of the analysis.

Table 2. Non-parametric Results for Population 1

	Chao1	ACES	Jackknife1	Jackknife2	Actual
Species Estimate	756	586	680	822	3,223
Error	77 %	82 %	79 %	74 %	

Although the non-parametric estimators seem to agree reasonably with each other about their predictions, all greatly underestimate the actual number of species present in this particular population. The standard deviation associated with the Chao1 estimator for this population is 54.8. In this instance, the output of the Chao1 method is over 44 standard deviations away from the truth. Tchebysheff's theorem can be applied to the Chao1 method output. The equation used for Tchebysheff's theorem, in regard to a point estimator, is:

$$P\left(\left|\hat{\theta}-\theta\right| < k\,\sigma_{\hat{\theta}_n}\right) \geq 1-\frac{1}{k^2} \tag{11}$$

where $E(\hat{\theta}_n)=\theta$ and $\sigma_{\hat{\theta}_n}=\sqrt{V(\hat{\theta}_n)}$, the standard deviation of the estimator. Furthermore, it states that 15/16 of the data falls within four standard deviations of the expected value. Applying this information to equation (11) yields:

$$P\left(537 < \hat{\theta} < 975\right) \geq \frac{15}{16} \tag{12}$$

Clearly, the actual number of species for this population does not fall within the above interval.

Parametrically fitting the sample data to a curve resulted in the following top-ranked distributions fitting the curve: 1) Exponential (two-parameter), 2) Exponential (one-parameter), 3) Gamma (three-parameter), 4) Weibull, and 5) Weibull (three-

parameter). Figures 5-9 depict the fits of each distribution to the data, where the *x*-axis represents the number of unique species found in the sample.

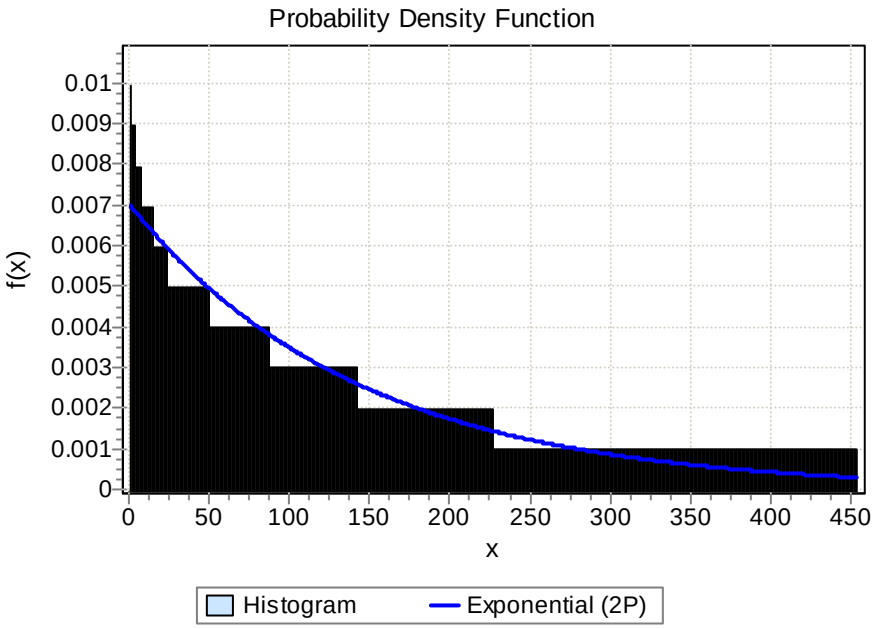

Figure 5. Exponential (two-parameter) Fit (Population 1)

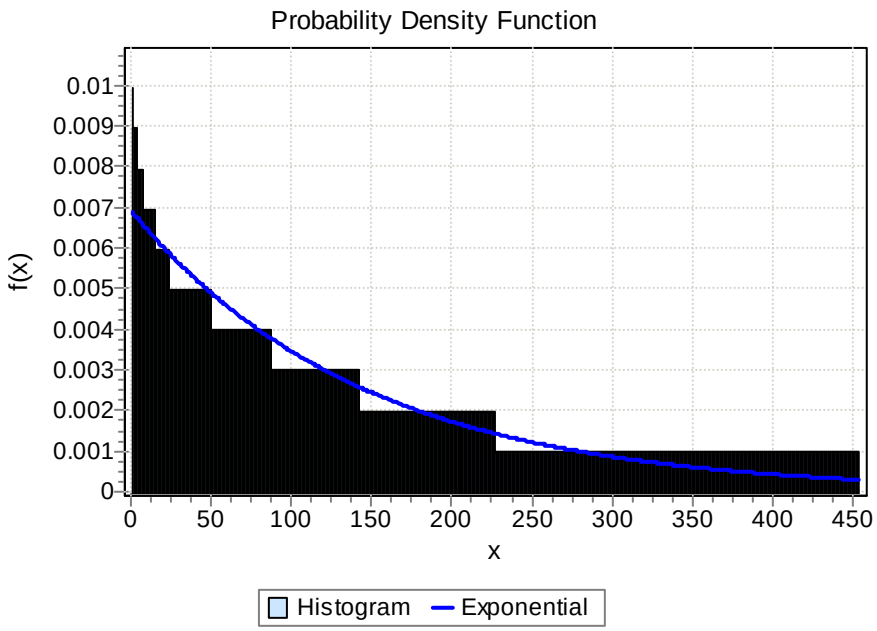

Figure 6. Exponential Fit (Population 1)

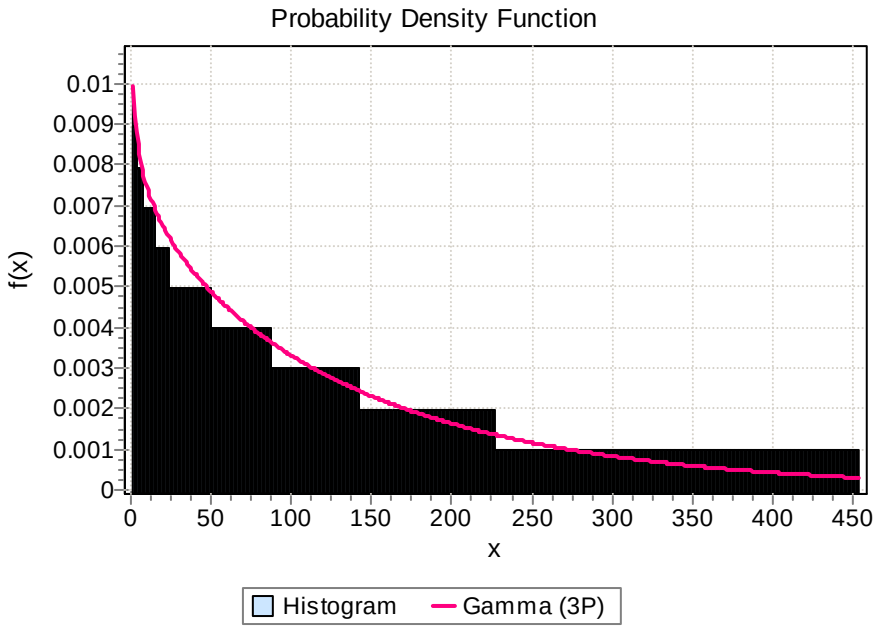

Figure 7. Gamma (three-parameter) Fit (Population 1)

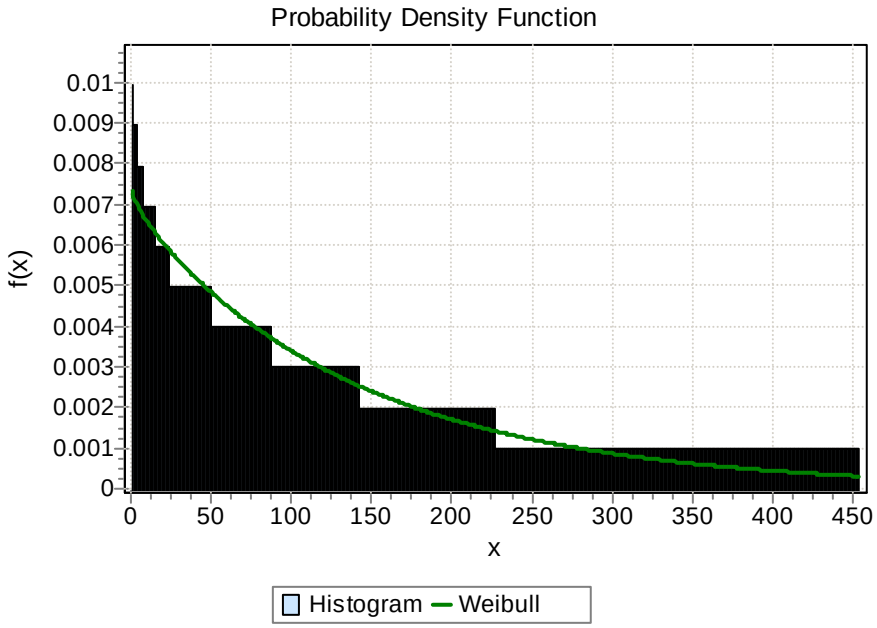

Figure 8. Weibull Fit (Population 1)

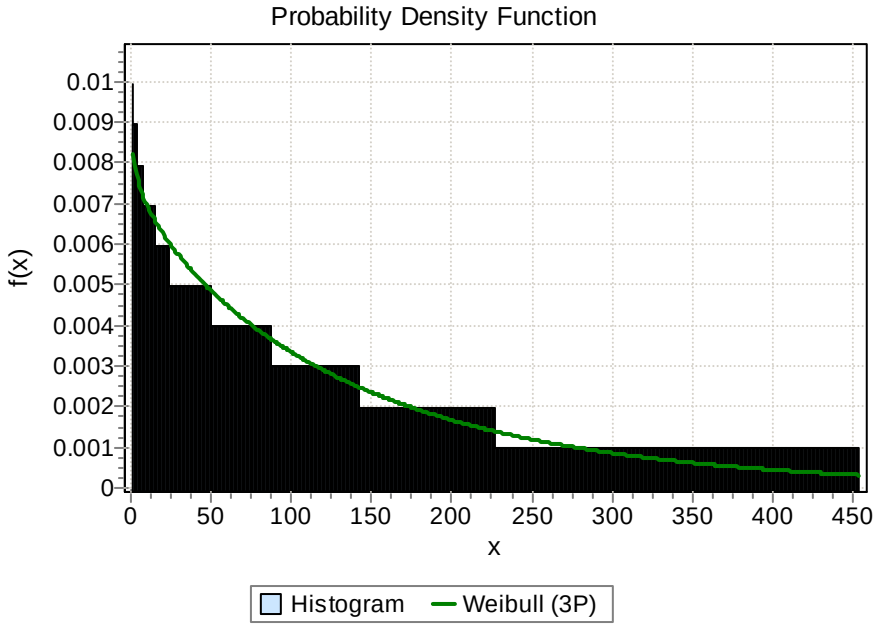

Figure 9. Weibull (three-parameter) Fit (Population 1)

The parameters and equations associated with each of the distributions can be found in Appendix D of this thesis. Table 3 contains the analysis results for Parametric fitting of Population 1.

Table 3. Parametric Results for Population 1

	Weibull	Weibull (3p)	Exponential (2p)	Exponential	Gamma (3p)	Actual
Species Estimate	2,500	2,663	2,351	2,362	2,518	3,223
Error	22 %	17 %	27 %	27 %	22 %	
Population Mean	146.67	150.45	143.36	144.18	139.27	
Population Standard Deviaton	148.43	160.23	142.36	144.18	152.95	

All the parametric curves estimated species composition closer to the truth than the non-parametric estimators, with parametric percent errors ranging between 17 and 27 percent and non-parametric percent errors ranging between 74 and 82 percent.

The three-parameter lognormal distribution was also analyzed (although not formally presented with this population), since this is one of the distributions which (according to the literature) are thought to describe the natural microbial population. It resulted in an over-estimation of the actual number of species, with a predicted value of 9,001 unique species.

Non-parametric Sample Size Versus Parametric Population Size Study

This subsection addresses the important issue of the variant nature of non-parametric sample sizes as well as the variant nature of parametric population sizes.

Population 1 was used to show that the non-parametric estimators will vary with a chosen sample size. This was done by choosing 15 random samples of differing sizes and applying the four non-parametric methods to the samples. Table 4 summarizes the results of the application of the non-parametric methods to the samples.

Table 4. Variability of Non-parametric Estimates Based on Sample Size

Sample Size	Chao1 Estimate	ACES Estimate	Jackknife1 Estimate	Jackknife2 Estimate	Actual
	Error	Error	Error	Error	
50	370	392	90	130	3,223
	89 %	88 %	97 %	96 %	
75	400	431	131	187	3,223
	88 %	87 %	96 %	94 %	
100	546	445	168	240	3,223
	83 %	86 %	95 %	99 %	
250	746	538	354	489	3,223
	77 %	83 %	89 %	85 %	
500	601	506	503	632	3,223
	81 %	84 %	84 %	80 %	
750	621	515	577	694	3,223
	81 %	84 %	82 %	78 %	
1,000	756	586	680	822	3,223
	77 %	82 %	79 %	74 %	
1,500	647	562	670	745	3,223
	80 %	83 %	79 %	77 %	
2,500	840	688	823	933	3,223
	74 %	79 %	74 %	71 %	
5,000	1,022	848	911	1,111	3,223
	68 %	74 %	72 %	66 %	
10,000	1,086	961	1,092	1,187	3,223
	66 %	70 %	66 %	63 %	
25,000	1,301	1,146	1,281	1,391	3,223
	60 %	64 %	60 %	57 %	
50,000	1,325	1,229	1,340	1,410	3,223
	59 %	62 %	58 %	56 %	
75,000	1,415	1,334	1,454	1,517	3,223
	56 %	59 %	55 %	53 %	
100,000	1,539	1,419	1,550	1,641	3,223
	52 %	56 %	52 %	49 %	

As noted above, the non-parametric estimators became more accurate as the sample size increased. The total population size for this study was just under a billion. A sample size of 100,000 was still unable to produce an estimate less than 49 percent away

from the actual number of species comprising this population. The following figure shows a graphically depiction of what the Chao1 method is estimating in relation to the sample size.

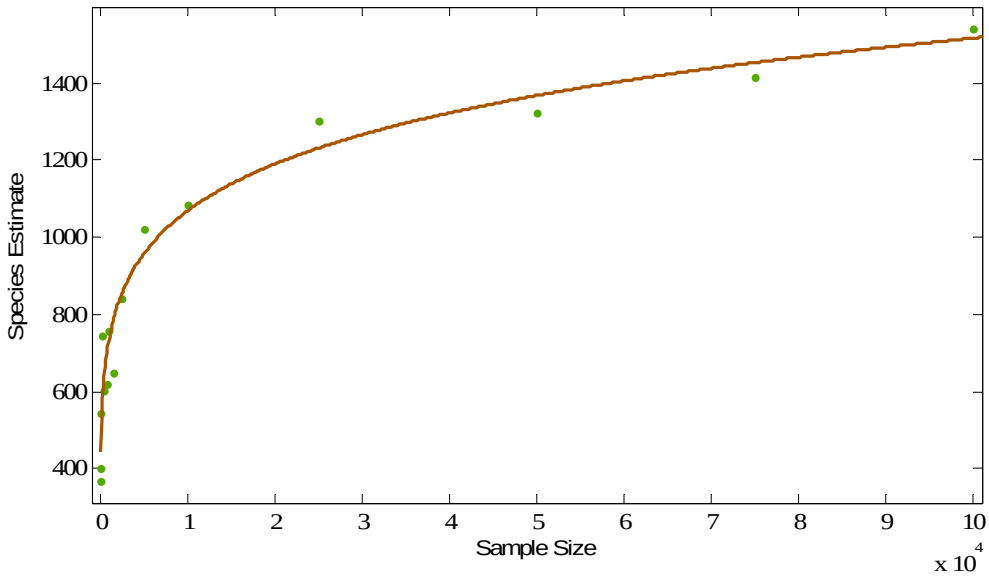

Figure 10. Chao1 Estimates

The following figure shows a graphically depiction of what the ACES method is estimating in relation to the sample size.

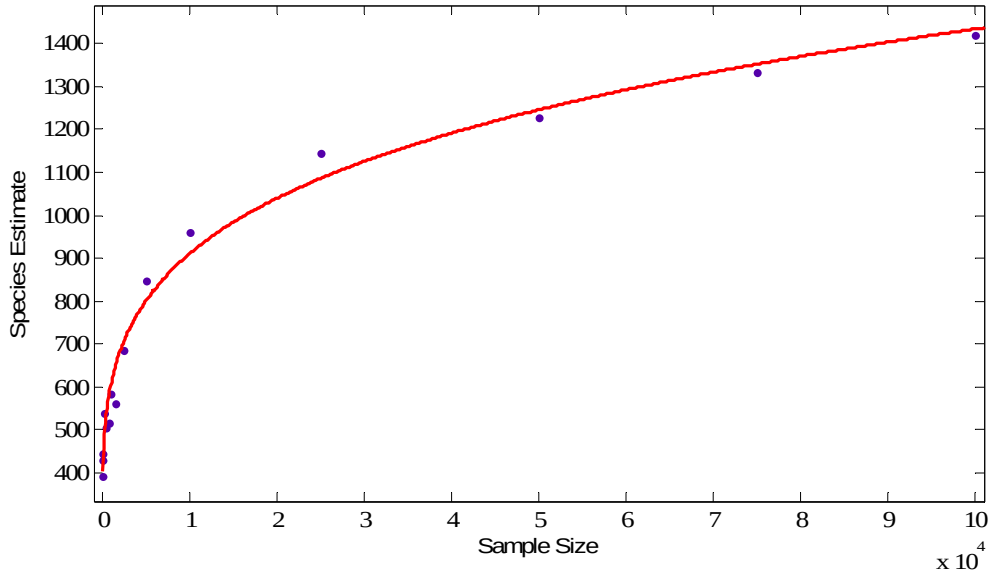

Figure 11. ACES Estimates

The graphs show both the Chao1 and the ACES methods providing improved estimates with an increased sample size. Consequently, a question to consider is then, what sample size would be necessary to result in an accurate estimate of the number of species. To address this issue, consider the following. For this particular population, the true number of species is $N = 3,223$. Of these species, 219 are species comprised of exactly 1 individual (f_1). Additionally, 102 are species comprised of exactly 2 individuals (f_2). Consider that all singletons and doubletons have been identified. Substituting this information into the Chao1 estimator given by equation (1) produces the following:

$$3223 = S + (219)^2 / (2*102) \qquad (13)$$

Solving for S gives that 2,988 unique species would be needed in addition to all 219 singletons and 102 doubletons to achieve the actual number of species in the population.

Given that, it may seem relatively easy to obtain a sample large enough to produce that number of unique species. However, for this population, as the sample size increases, the proportion of unique species and the proportion of singletons per sample decrease. Figures 12 and 13 depict this phenomenon graphically.

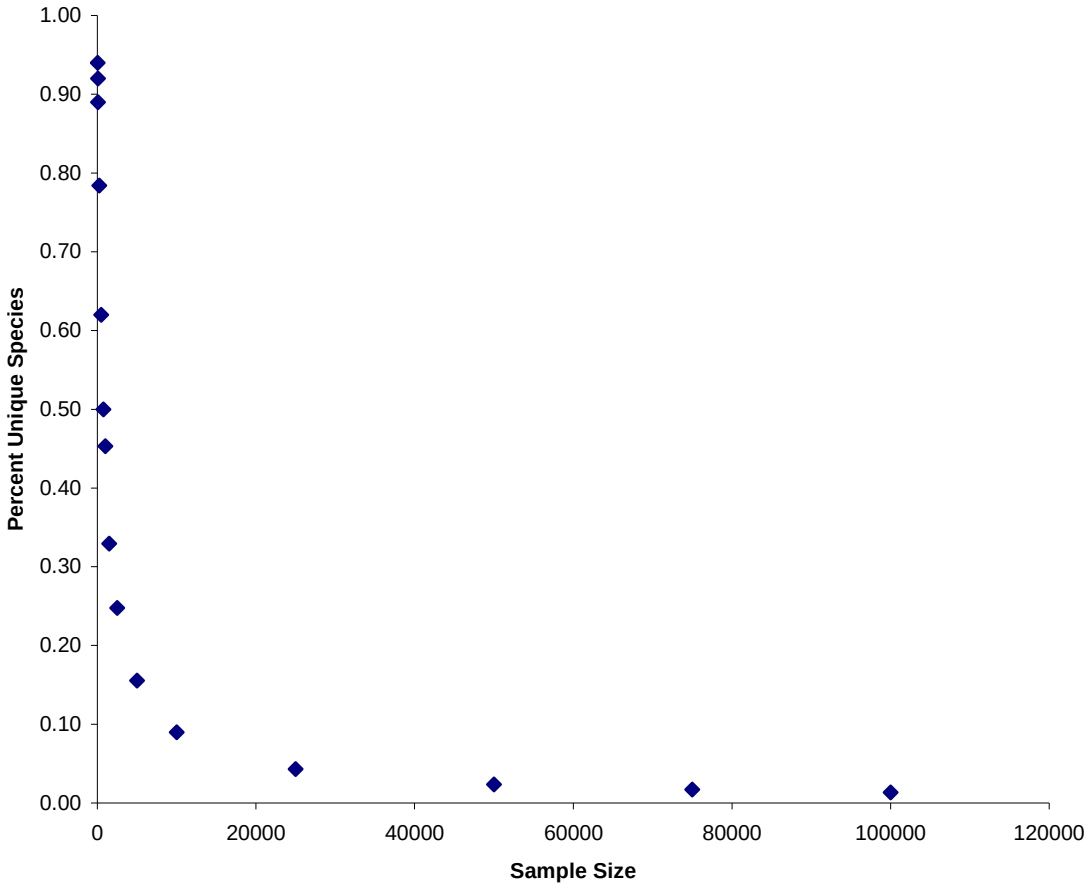

Figure 12. Percentage of Unique Species per Sample

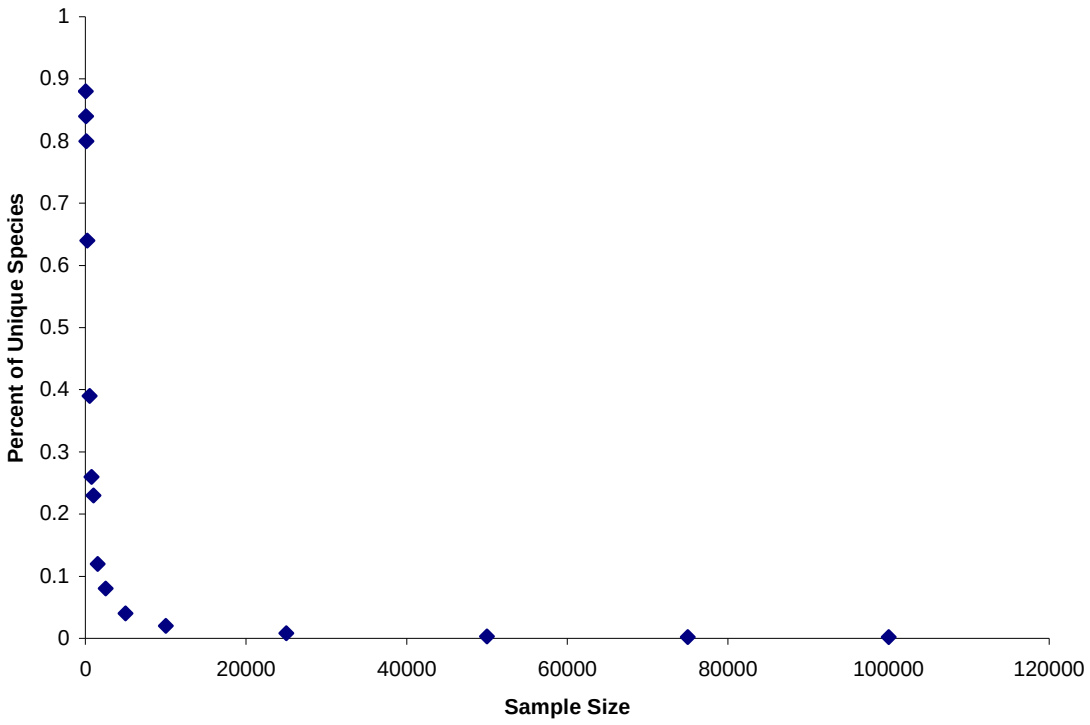

Figure 13. Percentage of Singletons per Sample

These graphs suggest that even as the sample size increases, the number of singletons and the number of unique species found are not increasing at fast enough rates to have a significant impact the accuracy of the non-parametric estimators. This leads to a new problem of having to exhaustively enumerate the population.

Additionally, with the methods currently available for microbial sampling, it is most likely impractical to have such a large sample size. However, as techniques for acquiring microbial samples increases, samples of sizes such as these may attainable. Furthermore, it may be that even increasing the sample size may not have the desired effect of increasing the accuracy of the non-parametric estimators.

While parametric estimators will not have the same variability as seen with the non-parametric estimators, the usage of such methods will depend on the population size used for the extrapolation. The following table summarizes the population sizes used for this study as well as the estimates produced.

Table 5. Parametric Estimates for Various Population Sizes

population size	species estimate
1.0×10^{6}	1,370
1.5×10^{6}	1,429
2.5×10^{6}	1,502
5.0×10^{6}	1,602
7.5×10^{6}	1,660
1.0×10^{7}	1,701
1.5×10^{7}	1,759
2.5×10^{7}	1,833
5.0×10^{7}	1,932
7.5×10^{7}	1,991
1.0×10^{8}	2,032
1.5×10^{8}	2,090
2.5×10^{8}	2,164
5.0×10^{8}	2,263
7.5×10^{8}	2,322
1.0×10^{9}	2,363
$1.5 \times 10^{9.}$	2,421
2.5×10^{9}	2,495
5.0×10^{9}	2,594
7.5×10^{9}	2,652
1.0×10^{10}	2,649
2.5×10^{10}	2,852
5.0×10^{10}	2,925
1.0×10^{11}	3,025
5.0×10^{11}	3,255
9.0×10^{11}	3,340
1.0×10^{12}	3,355
5.0×10^{12}	3,587
9.0×10^{12}	3,671

Figure 14 depicts the change in the estimates of a parametrically-fitted exponential curve, using the parameter estimate produce by this population's sample.

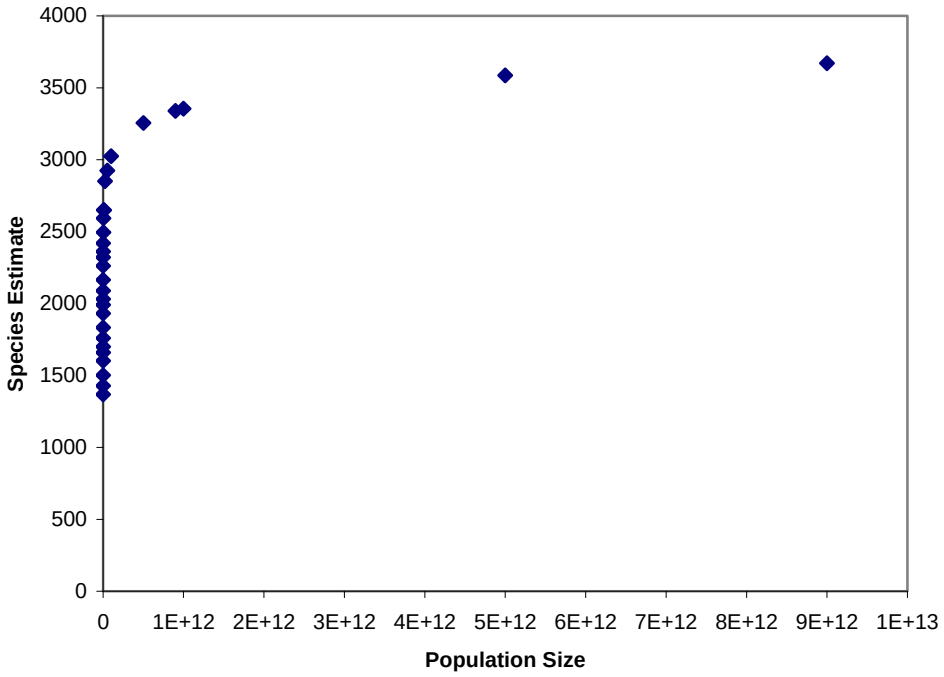

Figure 14. Parametric Estimates for Various Population Sizes

It is interesting to note that although the population size may increase by orders of magnitude, the estimate for the number of species in the population does not do the same. If this were the case, the graph would depict a more linear function. For this population, this does not appear to be the case. This implies that although the curve may not produce an asymptote, the rate at which the estimates increase will become increasingly slow. For example, the percent increase between a population size of 5×10^{11} and 9×10^{12} is approximately 1,700 percent, essentially, an increase by an order of magnitude. However, the species estimates of 3,255 and 3,671 (respectively) show a percent increase

of approximately 13 percent. Therefore, at some point, the population size will no longer have a drastic impact on the species estimate.

Population 2

Population 2 was created by using a gamma probability distribution with an alpha parameter value of 0.01 and a beta parameter value of 1000. This particular population was chosen to be smaller in size to see if perhaps a smaller population would provide better results for the non-parametric methods. The resulting population contained 532 individuals belonging to 167 different species. A random sample of size 100 was drawn and categorized as described in Chapter 3 of this document. Table 6 contains the pertinent population characteristics.

Table 6. Population 2 Specifications

Probability Distribution	Alpha Parameter	Beta Parameter	Population Size	Number of Species	Sample Size
Gamma	.01	1,000	532	167	100

The sample resulted in 38 unique species, of which 16 were singleton species. Figure 15 depicts a graphic representation of the sample.

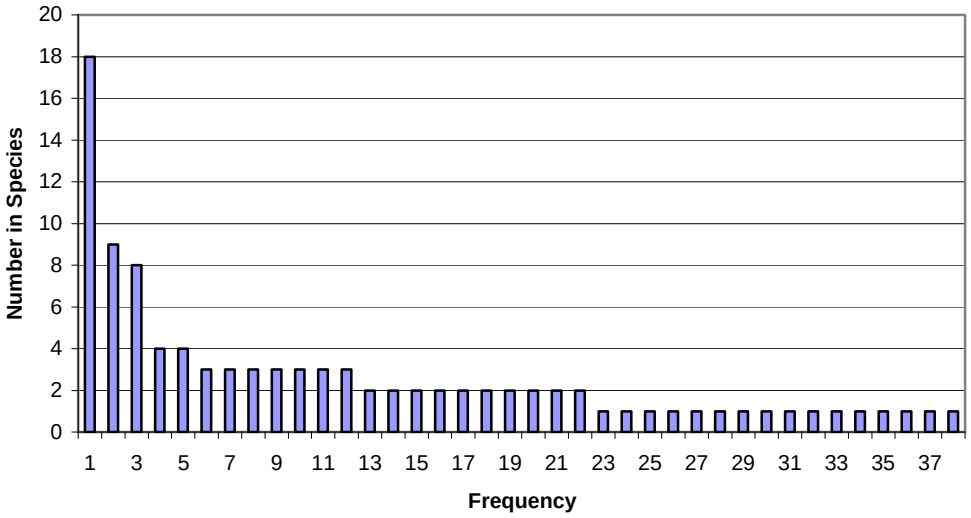

Figure 15. Sample Data Population 2

The four non-parametric methods were applied to the population data. Table 7 contains the results of the analysis.

Table 7. Non-parametric Results for Population 2

	Chao1	ACES	Jackknife1	Jackknife2	Actual
Species Estimate	51	47	54	60	167
Error	69 %	72 %	68 %	64 %	

Though the non-parametric estimators seem to reasonably agree with each other, it is apparent by looking at the above table that each method greatly underestimates the truth. The Chao1 method, for this population, produced a standard deviation of 8.37. In

this particular instance, the output of the Chao1 method is over 13 standard deviations away from the truth. Tchebysheff's theorem can be applied to the Chao1 method estimate for population 2. Applying the information from population 2 to equation (11) yields:

$$P\left(18 < \hat{\theta} < 84\right) \geq \frac{15}{16} \tag{14}$$

Clearly, the actual number of species for this population does not fall within the above interval.

Parametrically fitting the sample data to a curve resulted in the following top-ranked distributions fitting the curve: 1) Exponential, 2) Lognormal, 3) Weibull, 4) Gamma (three-parameter), and 5) Gamma. Figures 16-20 depict the fitting of the distributions to the data, where the *x*-axis represents the number of unique species found in the sample.

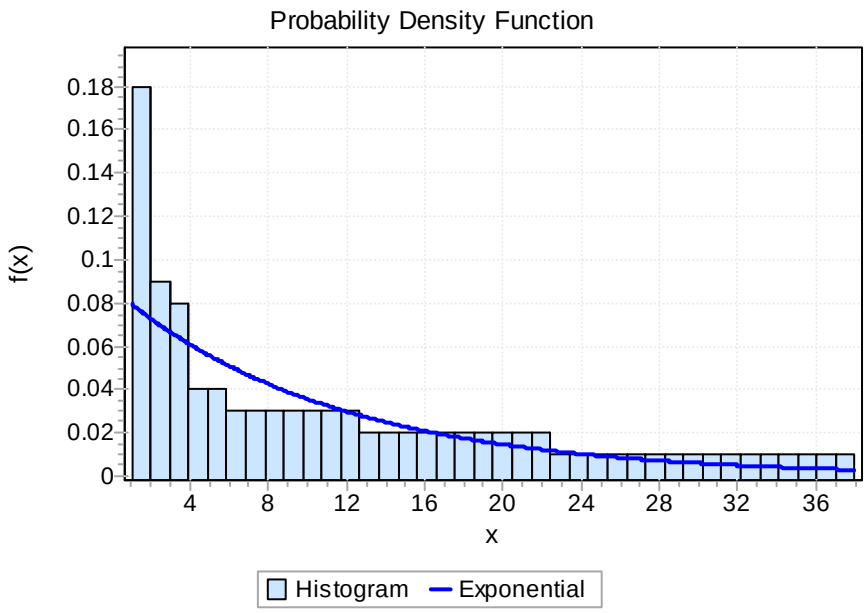

Figure 16. Exponential Fit (Population 2)

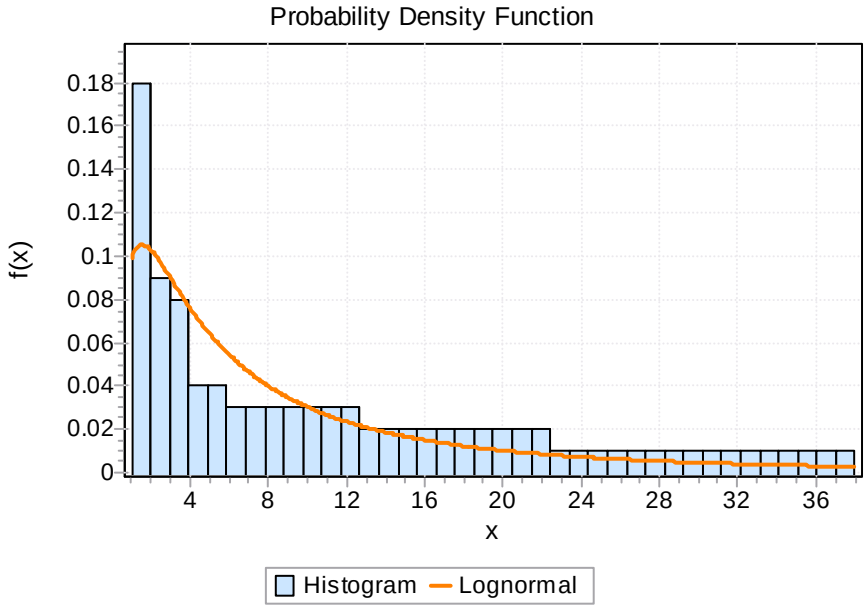

Figure 17. Lognormal Fit (Population 2)

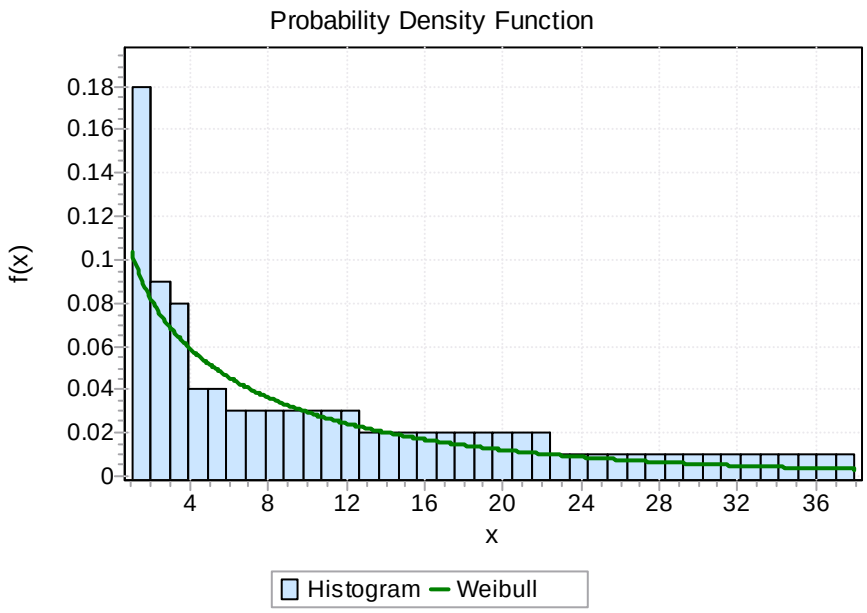

Figure 18. Weibull Fit (Population 2)

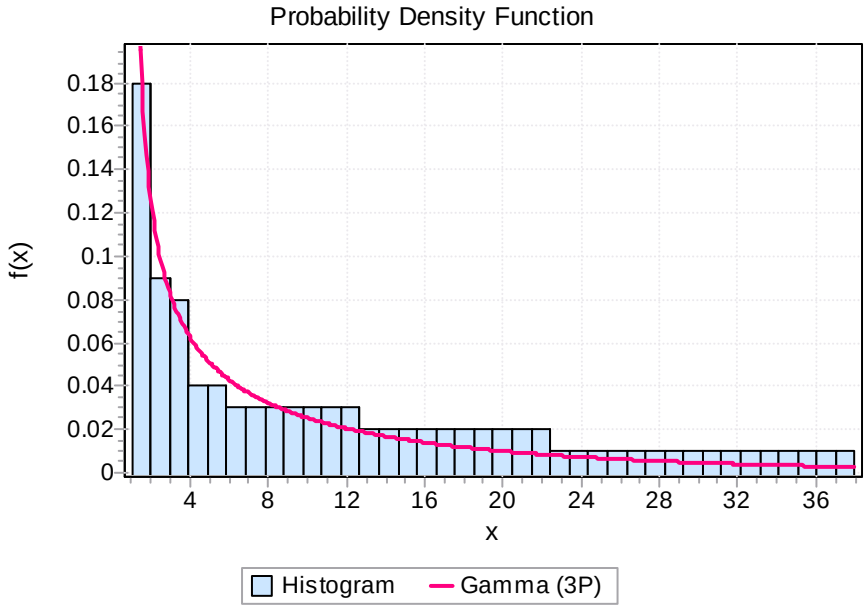

Figure 19. Gamma (three-parameter) Fit (Population 2)

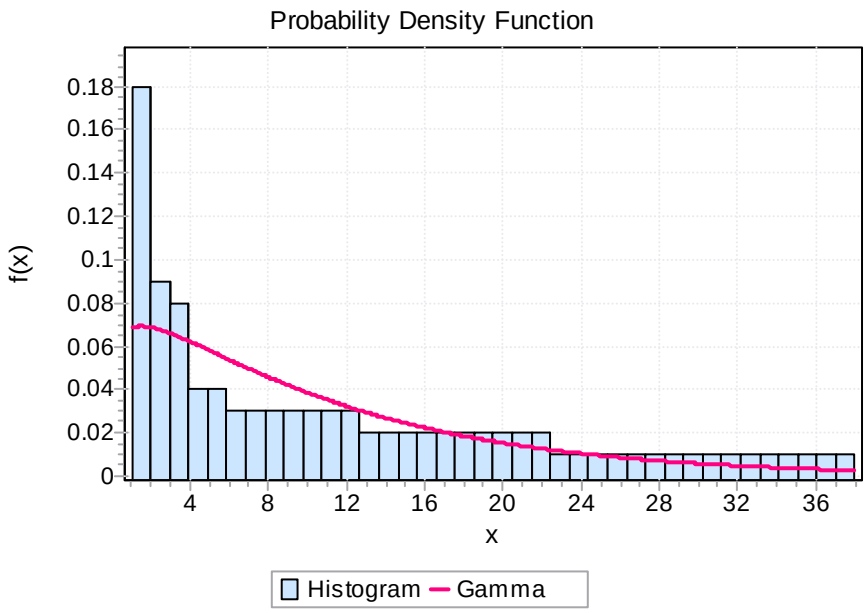

Figure 20. Gamma Fit (Population 2)

The parameters and equations associated with each of the distributions can be found in Appendix D of this thesis. Table 8 contains the analysis results for Parametric fitting of Population 2.

Table 8. Parametric Results for Population 2

	Exponential	Lognormal	Weibull	Gamma (3p)	Gamma	Actual
Species Estimate	51	59	56	54	48	167
Error	69 %	65 %	66 %	68 %	71 %	
Population Mean	11.188	10.46	10.428	6.3227	11.088	
Sample Standard Deviation	11.188	14.644	11.113	9.1057	14.644	

As noted above, all the parametric curves estimated species composition in the same range as the non-parametric methods, with raw estimates for the non-parametric methods ranging between 51 and 60. Raw estimates for the parametric methods resulted in a range between 48 and 59.

For this small population, the use of the parametric methods proved no closer to the truth than the non-parametric methods. That is to say, both categories of methods performed poorly, with a non-parametric percent error average and a parametric percent error average of 68 percent.

Population 3

Population 3 was created from a gamma distribution with an alpha parameter value of 0.001 and a beta parameter value of 500. This particular population was chosen to be smaller in size than population 1, but larger in size than population 2. The resulting population contained 6,197,020 individuals belonging to 3,218 different species. A random sample of size 2,841 was drawn (to reflect a sample size similar to that drawn from the wetlands soil) and categorized as described in Chapter 3 of this document. Table 9 contains the pertinent population characteristics.

Table 9. Population 3 Specifications

Probability Distribution	Alpha Parameter	Beta Parameter	Population Size	Number of Species	Sample Size
Gamma	.001	500	6,197,020	3,218	2,841

The sample resulted in 404 unique species, of which 182 were singleton species. Figure 21 depicts a graphic representation of the sample.

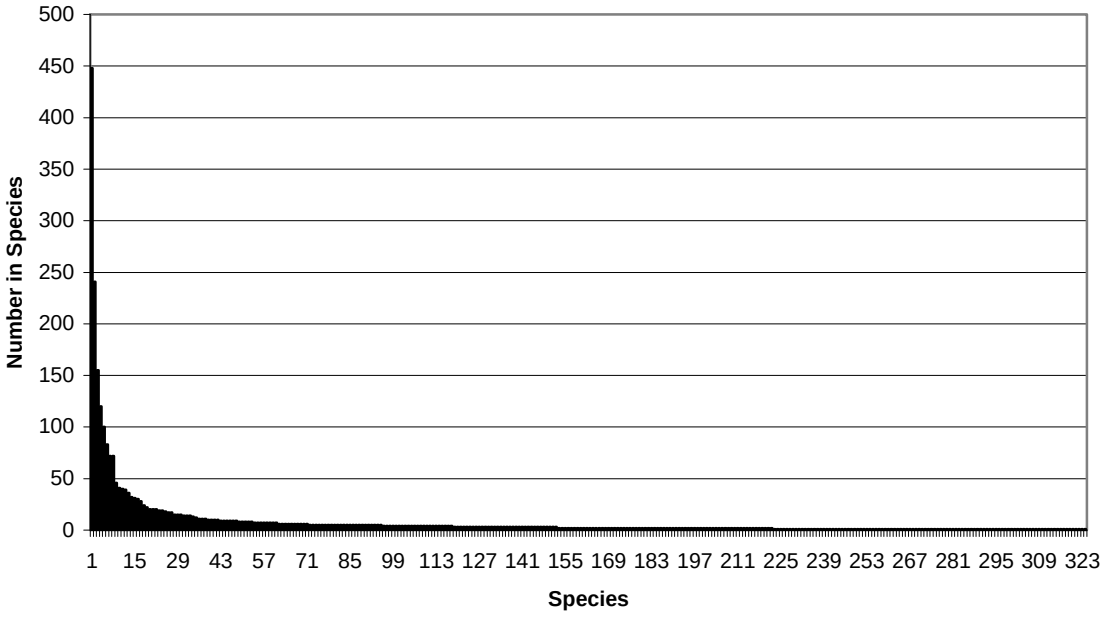

Figure 21. Sample Data Population 3

The four non-parametric methods were applied to the population data. Table 10 contains the results of the analysis.

Table 10. Non-parametric Results for Population 3

	Chao1	ACES	Jackknife1	Jackknife2	Actual
Species Estimate	641	499	586	698	3,218
Error	80 %	84 %	82 %	78 %	

The non-parametric estimators seem to reasonably agree with each other, it is apparent by looking at the above table that each method greatly underestimates the truth. For this population, the Chao1 method produced a standard deviation of 47.6. In this particular instance, the output of the Chao1 method is over 54 standard deviations away from the truth. Tchebysheff's theorem can be applied to the Chao1 method estimate for population 3. Applying the information from population 2 to equation (11) yields:

$$P\left(451 < \hat{\theta} < 831\right) \geq \frac{15}{16} \tag{15}$$

Clearly, the actual number of species for this population does not fall within the above interval.

Parametrically fitting the sample data to a curve resulted in the following top-ranked distributions fitting the curve: 1) Gamma (three-parameter), 2) Lognormal, 3) Weibull (three-parameter), 4) Weibull, and 5) Gamma. Figures 22-26 depict the fitting of the distributions to the sample data, where the *x*-axis represents the number of unique species found in the sample.

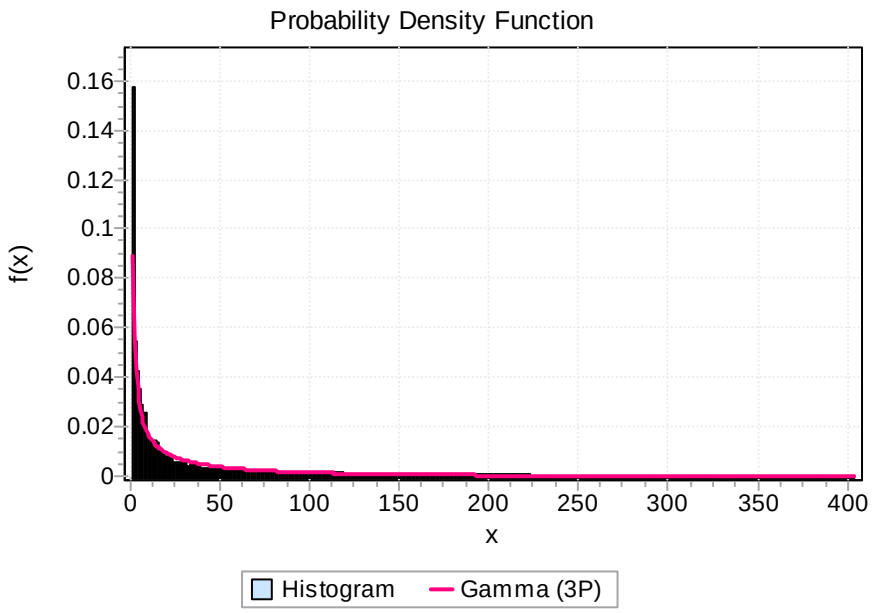

Figure 22. Gamma (three-parameter) Fit (Population 3)

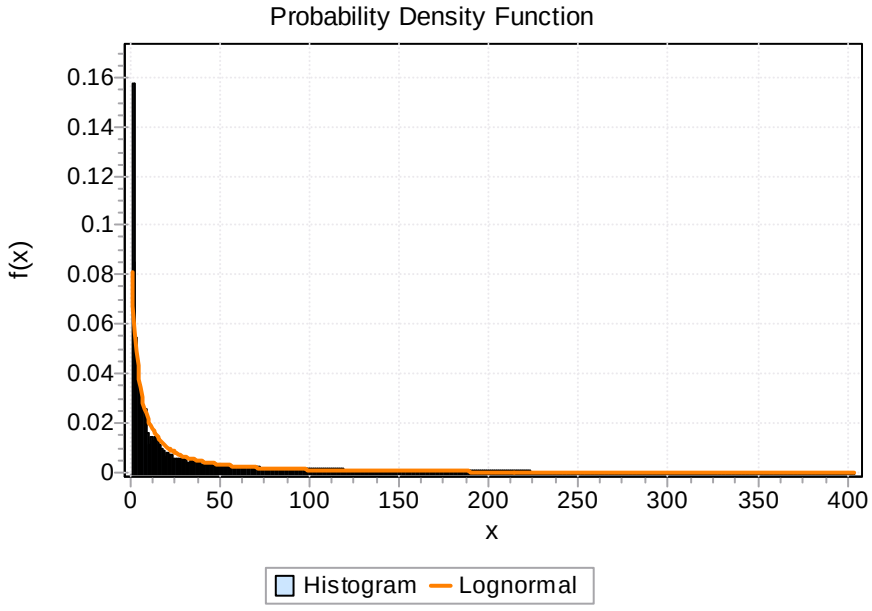

Figure 23. Lognormal Fit (Population 3)

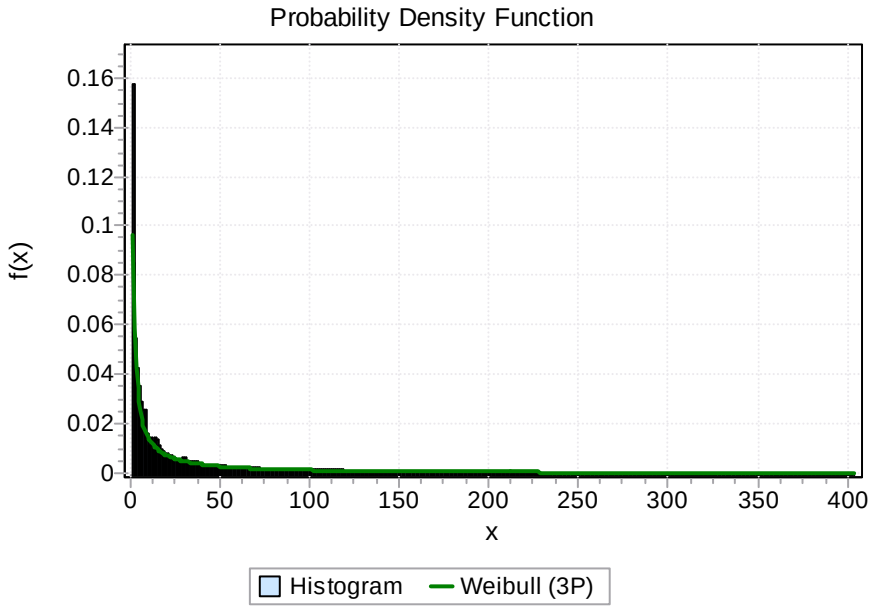

Figure 24. Weibull (three-parameter) Fit (Population 3)

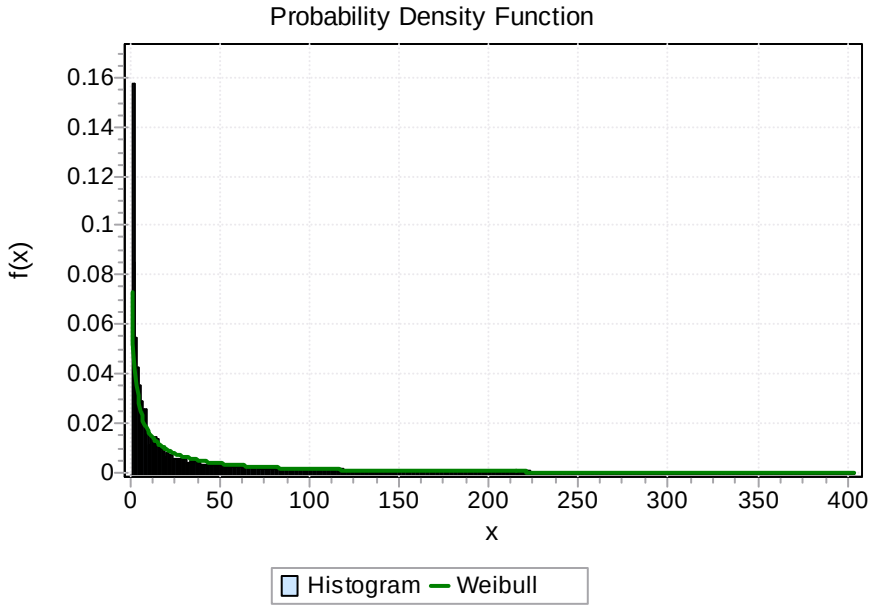

Figure 25. Weibull Fit (Population 3)

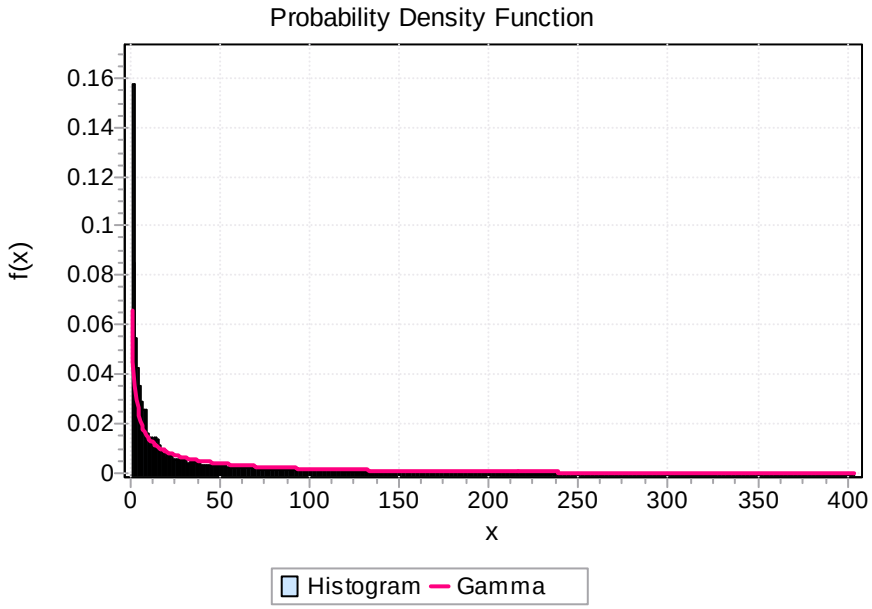

Figure 26. Gamma Fit (Population 3)

The parameters and equations associated with each of the distributions can be found in Appendix D of this thesis. Table 11 contains the analysis results for parametric fitting of Population 3.

Table 11. Parametric Results for Population 3

	Gamma (3p)	Lognormal	Weibull (3p)	Weibull	Gamma	Actual
Species Estimate	952	6,893	6,976	2,358	1,272	3,218
Error	70 %	114 %	117 %	27 %	60 %	
Population Mean	39.843	72.345	56.569	61.651	60.486	
Sample Standard Deviation	51.854	268.49	130.04	112.84	88.824	

It is interesting to note that for this particular population, there were two parametric distributions which overestimated the number of species comprising the population. Of the three which did not provide an overestimation, the outputs of all three were closer to the truth than any of the non-parametric methods used.

It appears that for this particular population, the use of the parametric methods proved closer to the truth than the non-parametric methods.

Population 4

Population 4 was created by using an alpha parameter value of 0.001 and a beta parameter value of 5,000. The resulting population contained 849,450 individuals belonging to 13,484 different species. This was chosen to be a more diverse population than population 3. A random sample of size 2,840 was drawn (to reflect a sample size similar to that drawn from the wetlands soil) and categorized as described in Chapter 3 of this document. Table 12 contains the pertinent population characteristics.

Table 12. Population 4 Specifications

Probability Distribution	Alpha Parameter	Beta Parameter	Population Size	Number of Species	Sample Size
Gamma	.001	5,000	849,450	13,484	2,840

The sample resulted in 929 unique species, of which 639 were singleton species. Figure 27 depicts a graphic representation of the sample.

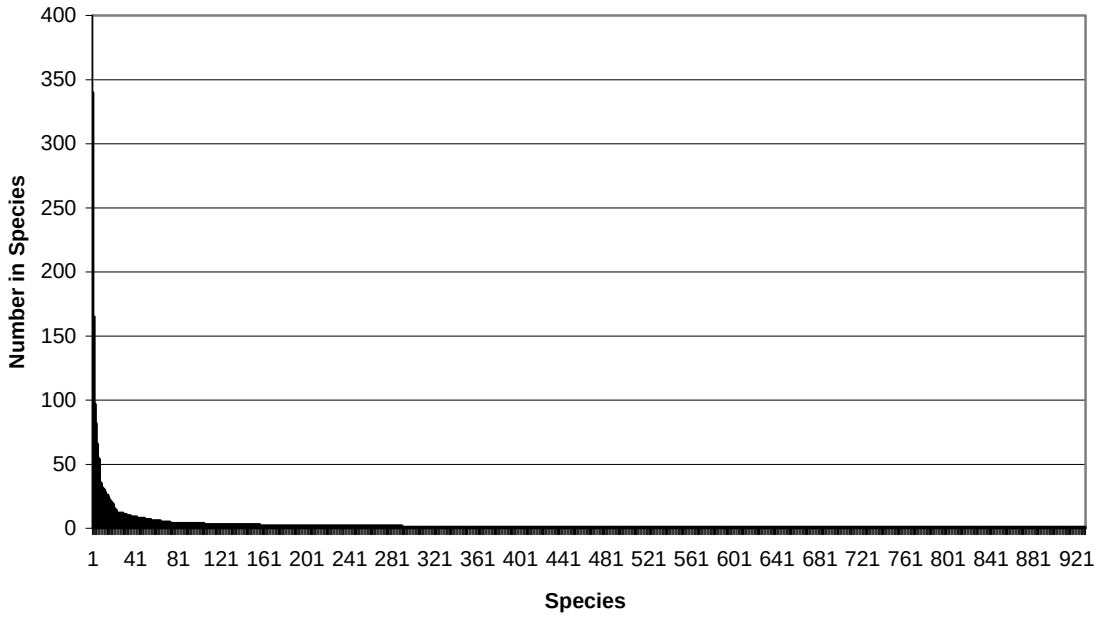

Figure 27. Sample Data Population 4

The four non-parametric methods were applied to the population data. Table 13 contains the results of the analysis.

Table 13. Non-parametric Results for Population 4

	Chao1	ACES	Jackknife1	Jackknife2	Actual
Species Estimate	2,464	1,611	1,568	2,073	13,484
Error	82 %	88 %	88 %	85 %	

Although the non-parametric estimators seem to agree reasonably with each other about their predictions, it is apparent by looking at the above table that all the methods are a great deal smaller than the truth. The Chao1 method has an associated equation for

a variance, with a resultant standard deviation of 184.4. In this particular instance, the output of the Chao1 method is over 59 standard deviations away from the truth. Tchebysheff's theorem can be applied to the Chao1 method estimate for population 4. Applying the information from population 4 to equation (11) yields:

$$P\left(1,728 < \hat{\theta} < 3,200\right) \geq \frac{15}{16} \qquad (16)$$

Clearly, the actual number of species for this population does not fall within the above interval.

Parametrically fitting the sample data to a curve resulted in the following top-ranked distributions fitting the curve: 1) Gamma (three-parameter), 2) Lognormal, 3) Weibull, 4) Weibull (three-parameter) and, 5) Gamma. Figures 28-32 depict the fitting of the distributions to the sample data, where the *x*-axis represents the number of unique species found in the sample.

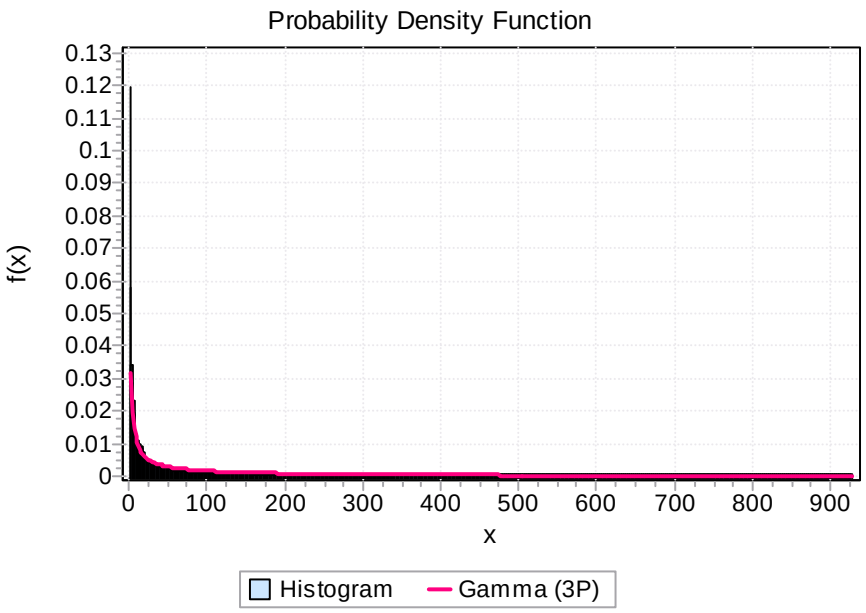

Figure 28. Gamma (three-parameter) Fit (Population 4)

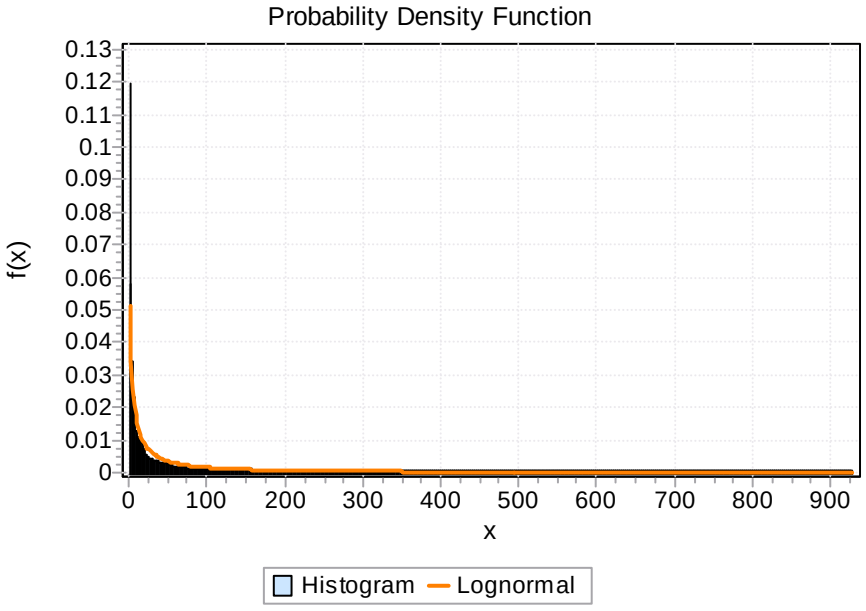

Figure 29. Lognormal Fit (Population 4)

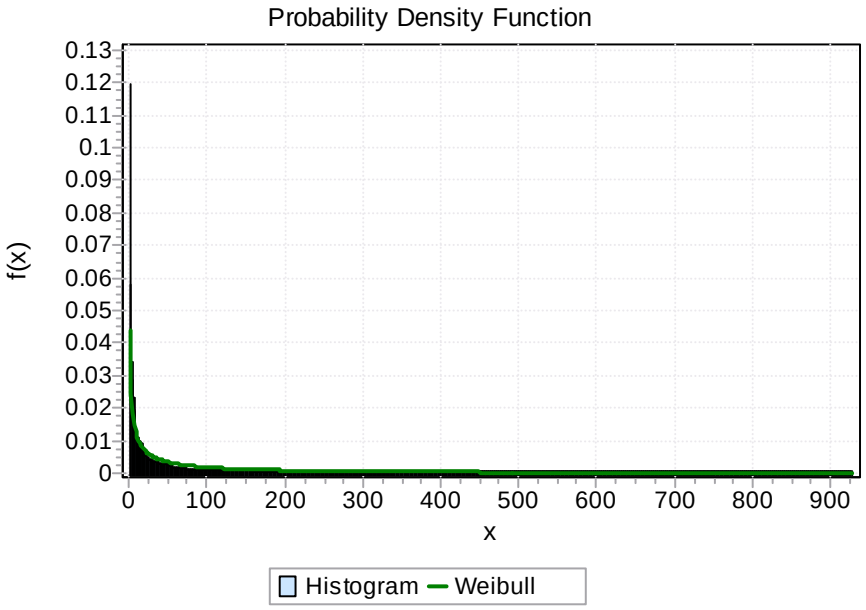

Figure 30. Weibull Fit (Population 4)

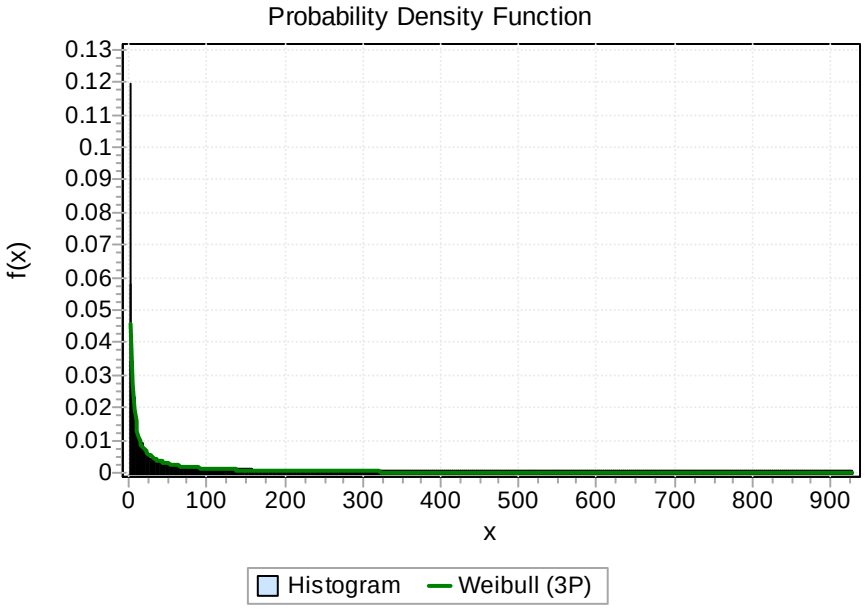

Figure 31. Weibull (three-parameter) Fit (Population 4)

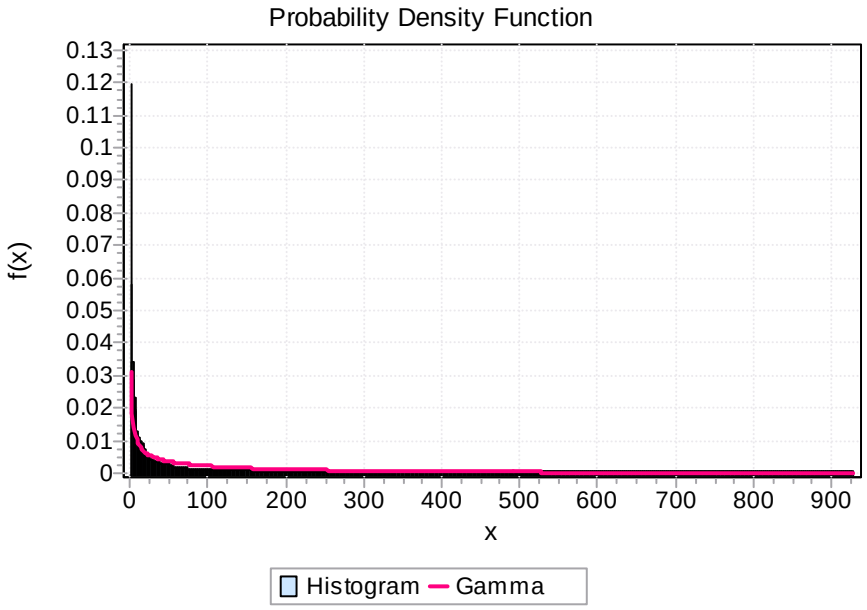

Figure 32. Gamma Fit (Population 4)

The parameters and equations associated with each of the distributions can be found in Appendix D of this thesis. Table 14 contains the results for the parametric fitting of Population 4.

Table 14. Parametric Results for Population 4

	Gamma (3p)	Lognormal	Weibull	Weibull (3p)	Gamma	Actual
Species Estimate	3,427	10,674	5,267	6,085	2,526	13,484
Error	75 %	21 %	61 %	55 %	81 %	
Population Mean	190.45	366.68	201.92	190	186.96	
Sample Standard Deviation	276.93	2524.0	317.98	353.2	257.14	

It is interesting to note with this population that the results produced by the fitting of the gamma distributions to the data closely resemble the results produced by the Chao1 method. Meanwhile, the Weibull and lognormal populations provided estimates closer to the truth, with the lognormal estimate producing a result approximately one standard deviation from the true parameter.

Wetlands Data

The final set of data which was analyzed was the sample data collected by Capt Elisabeth Leon from the constructed wetlands mesocosm. The mesocosms are columns of plants and soil which are representative of the constructed wetlands as a whole (Bishop; 2006:2). Therefore, any extrapolation was done with respect to the entire wetlands. The size of the wetlands is 120 feet by 60 feet by 4.5 feet. Converting these dimensions to centimeters and multiplying times the porosity of the soil leads to approximately 1.2×10^9 grams of soil in the wetlands. As discussed in Chapter 2, there are between 3×10^6 and
5×10^9 bacteria living in a gram of soil (Martin and Foch, 1977). This leads to a population size estimation of between 3.58×10^{15} and 5.97×10^{18} bacteria in the constructed wetlands. Figure 33 depicts the sample taken from the wetlands soil.

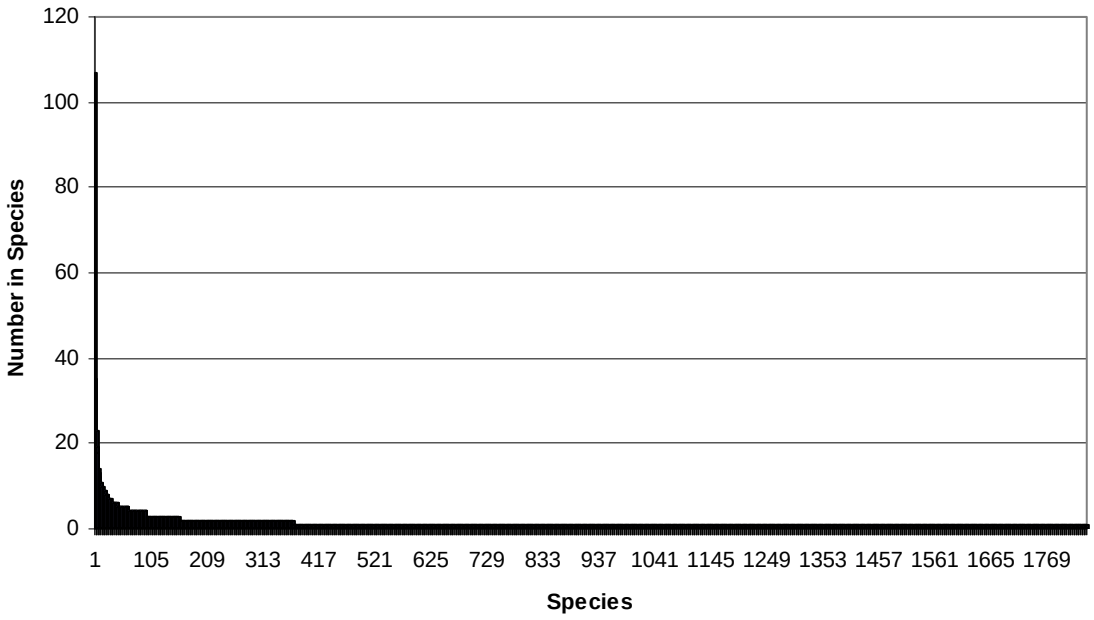

Figure 33. Wetlands Soil Sample

The sample size was comprised of 1,841 unique sequences. There were 1,474 singleton sequences and 211 doubleton sequences. It is interesting to note there was one species from which individuals were captured 107 times. The next largest group of species contained 35 sequences. The four non-parametric methods were applied to the population data. Table 15 contains the results of the analysis.

Table 15. Non-parametric Results for Wetlands Data

	Chao1	ACES	Jackknife1	Jackknife2	Actual
Species Estimate	6,990	4,399	3,315	4,577	Unknown

The standard deviation associated with the Chao1 method for this particular data set was 450.2. Since the truth is unknown, it is impossible to accurately assess how many (if any) standard deviations the Chao1 method is away from the truth. However, Tchebysheff's theorem can be applied to the Chao1 method estimate for the wetlands soil sample. Applying the information from this population to equation (11) yields:

$$P\left(5{,}190 < \hat{\theta} < 8{,}790\right) \geq \frac{15}{16} \qquad (17)$$

Then, according to this interval, the estimate for the actual number of species, would be, with a probability of 94 percent, between 5,190 and 8,790, though in no other population tested did the actual number of species fall into this respective interval.

Parametric fitting of the data resulted in the following top-ranked distribution fits: Gamma (three-parameter), Weibull, Exponential, Lognormal (three-parameter), and Exponential (two-parameter). Figures 34-38 depict the fitted distributions against the sample data, where the *x*-axis represents the number of unique species found in the sample.

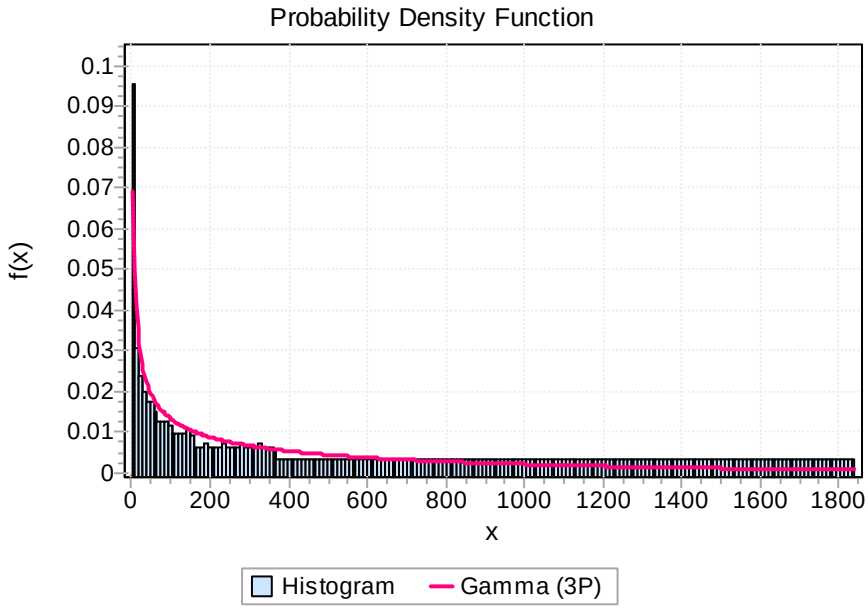

Figure 34. Gamma (three-parameter) Fit

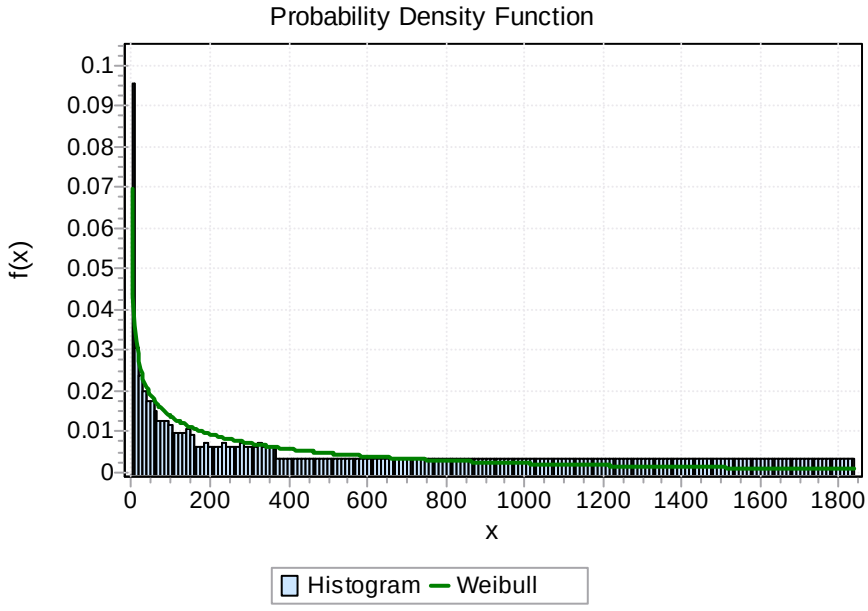

Figure 35. Weibull Fit

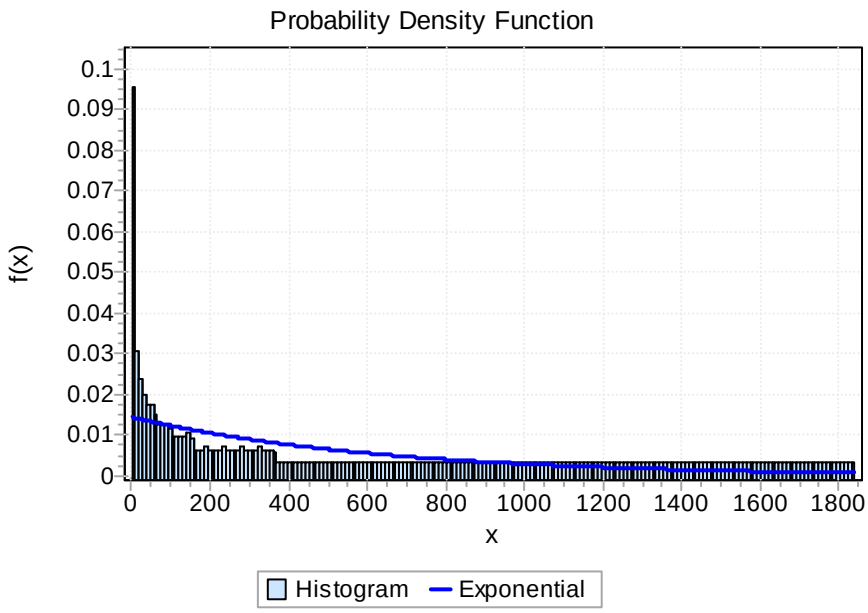

Figure 36. Exponential Fit

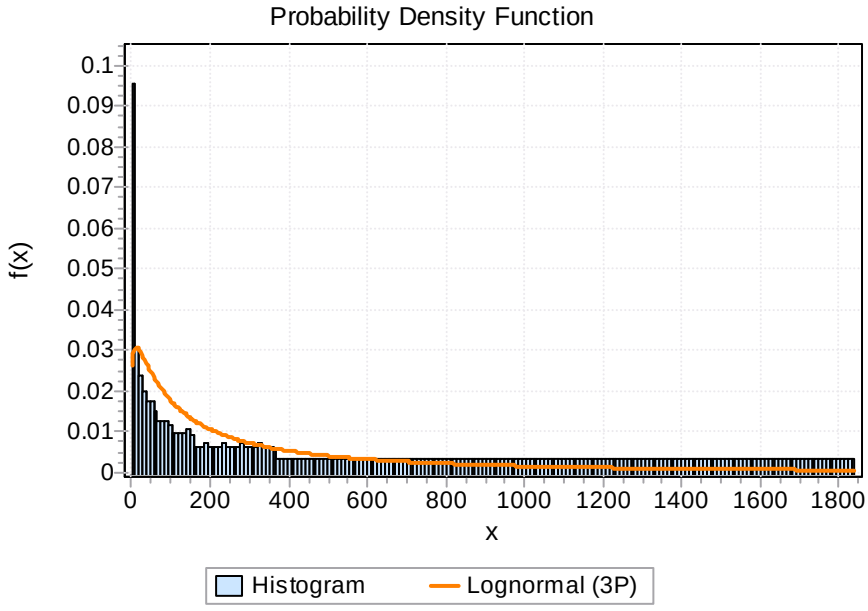

Figure 37. Lognormal (three-parameter)

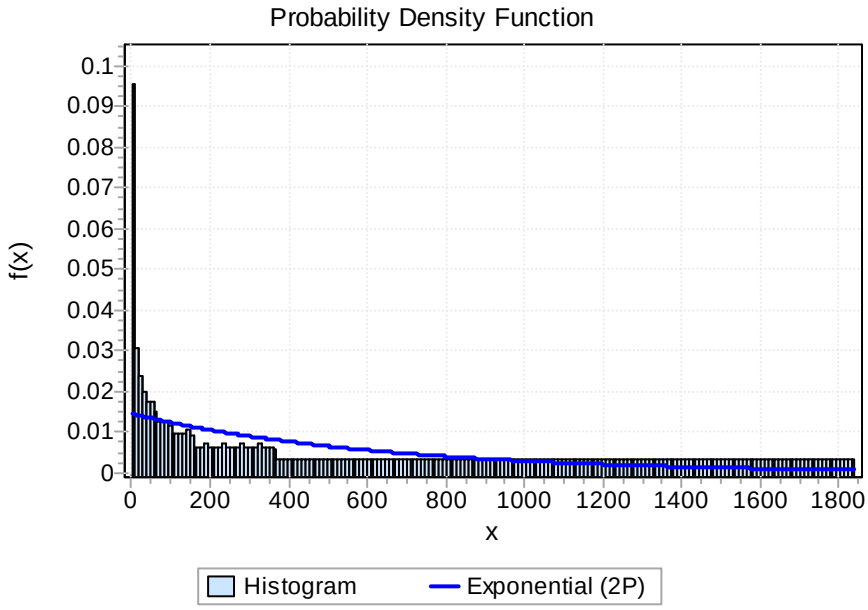

Figure 38. Exponential (two-parameter)

The parameters and equations associated with each of the distributions can be found in Appendix D of this thesis. Table 16 contains the analysis results for parametric fitting of the wetlands soil data.

Table 16. Parametric Results for Wetlands Soil Data

	Gamma (3p)	Weibull	Exponential	Exponential (2p)
Lower Confidence Interval Species Estimate	34,653	67,497	19,069	19,070
Upper Confidence Interval Species estimate	43,814	93,539	23,719	23,720

The species estimate for the lognormal curve is not included in the table above. The estimation for the number of species produced using a lognormal fit was producing a species estimate well over 900,000 for the lower confidence interval. This is an order of magnitude larger than the rest of the parametric estimates. As such, it may be that the lognormal is resulting in a overestimation of the number of species in this population.

The truth surrounding the true number of species in the soil is unknown. However, it is interesting to note the magnitude of differences between the non-parametric and parametric method applications to this data set. Table 17 compares both categories of methodology.

Table 17. Non-Parametric and Parametric Results for Wetlands Soil Data

Non-Parametric Methods					
	Chao1	ACES	Jackknife1	Jackknife2	Actual
Species Estimate	6,990	4,399	3,315	4,577	Unknown
Parametric Methods					
	Gamma (3p)	Weibull	Exponential	Exponential (2p)	Actual
Lower Confidence Interval Species Estimate	34,653	67,497	19,069	19,070	Unknown
Upper Confidence Interval Species estimate	43,814	93,539	23,719	23,720	Unknown

The sample size obtained was significant in size. It is large enough to give a good indication of what the underlying distribution could look like. As such, the results from the parametric fitting of the data should be considered as a better alternative to the non-parametric methods.

Wetlands Sub-Study

This subsection further addresses the issue of the variant nature of non-parametric sample sizes as well as the variant nature of parametric population sizes. The intention of this sub-study is to further emphasize the results previously obtained from the sub-study conducted on Population 1.

Unlike the previous sub-study, real-world data was utilized for this sub-study. Random samples, ranging in size from 200 to 2,820 points were drawn from the wetlands soil sample provided by Leon. Table 18 lists the results of applying the Chao1 and ACES non-parametric methods to the various sample sizes.

Table 18. Chao1 and ACES Non-Parametric Results for Various Sample Sizes

Sample Size	Chao1	ACES
200	2,223	1,020
500	3,445	1,960
750	2,958	1,871
1,000	3,412	2,364
1,500	4,959	3,072
2,000	5,770	3,563
2,820	6,990	4,399

Figure 39 depicts the results graphically, using a linear scale for the *x*-axis. Figure 40 depicts the same results using a log scale for the *x*-axis.

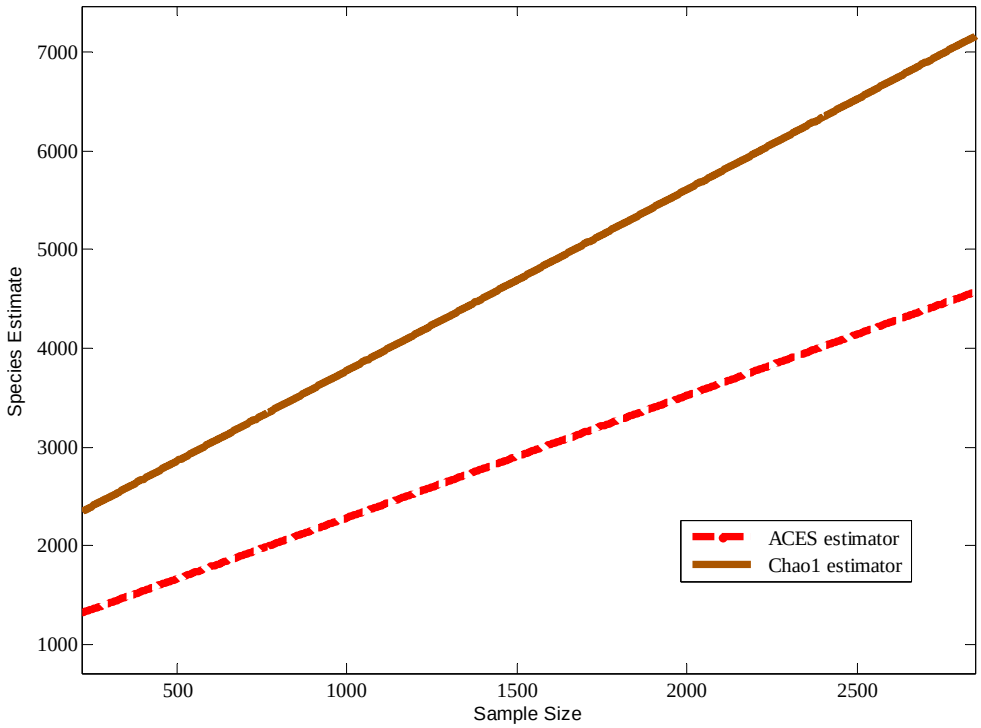

Figure 39. Chao1 and ACES Estimates for Various Sample Sizes

(Linear Scale)

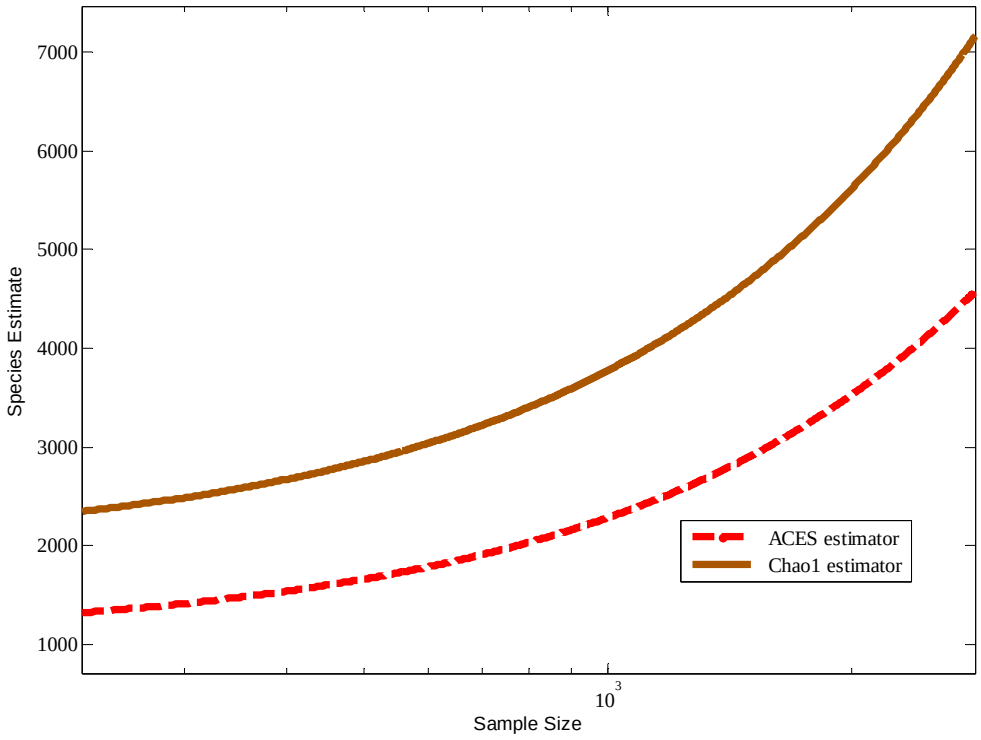

Figure 40. Chao1 and ACES Estimates for Various Sample Sizes (Log Scale)

Neither graph shows the data tending towards as an asymptote, as was the case with the results seen for the sub-study for Population 1. This may be indicative that the sample sizes are nowhere near large enough for the slowing of the species estimate that was observed earlier in this chapter. Using the information from Figure 39, approximations for the minimum sample sizes needed to obtain species estimates in the same range as the parametric estimators can be obtained. Table 19 summarizes the results.

Table 19. Minimum Sample Sizes Required to Equate Non-parametric and Parametric Estimation Methods

Parametric Method	Parametric Estimate Range	Chao1 Sample Size	ACES Sample Size
Gamma (3p)	34,653 - 43,814	17,770-22,830	27,140-34,530
Weibull	67,497-93,539	35,740-49,940	53,660-74,690
Exponential	19,069-23,719	9,340-11,870	14,550-18,310

It is imperative to convey that the information provided in Table 19 are estimates based on taking sub-samples of the wetlands soil sample data. A non-trivial assumption regarding the estimates is that the composition of the sample remains relatively constant, which may not be a valid assumption. Figure 40 indicates that on a log scale, the estimates are not linear in relation to the sample size. This may further indicate that the composition of the sample does not remain the same as the sample size increase. The estimates in Table 19 are provided for informational purposes only.

While the non-parametric estimators once again varied with the sample size of the population, the parametric estimators, although invariant to sample size, will vary with the population size. With a great deal of uncertainty surrounding the population size of the wetlands, observation of changes in the species estimation as the population size varies becomes important. As such, population sizes ranging between 1×10^{15} and 5×10^{21} were applied to the gamma (3p), Weibull, and exponential fits. Table 20 summarizes the results.

Table 20. Various Population Sizes for Parametric Fits

Population Size	Gamma (3p)	Weibull	Exponential
1.0×10^{15}	32,965	63,450	18,216
3.58×10^{15}	34,553	67,497	19,069
5.0×10^{15}	34,969	68,751	19,234
1.0×10^{16}	35,833	71,073	19,673
5.0×10^{16}	37,840	76,554	20,692
1.0×10^{17}	38,705	78,951	21,130
5.0×10^{17}	40,715	84,604	22,149
1.0×10^{18}	41,580	87,074	22,588
5.0×10^{18}	43,592	92,891	23,606
5.97×10^{18}	43,814	93,539	23,719
1.0×10^{19}	44,459	95,431	24,045
5.0×10^{19}	46,472	101,407	25,064
1.0×10^{20}	47,339	104,014	25,502
5.0×10^{20}	49,354	110,143	26,521
1.0×10^{21}	50,222	112,841	26,960
5.0×10^{21}	52,238	119,090	27,978

Figures 41 and 42 depict the information graphically, using a linear and log scale respectively for the *x*-axis.

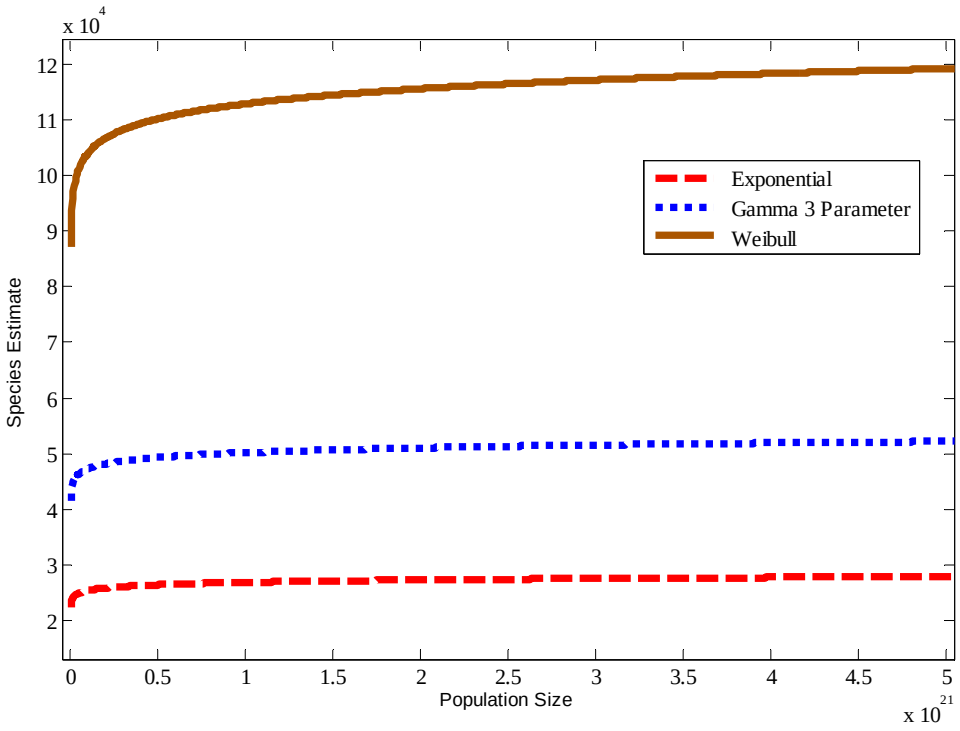

Figure 41. Parametric Estimates for Various Wetlands Population Sizes (Linear Scale)

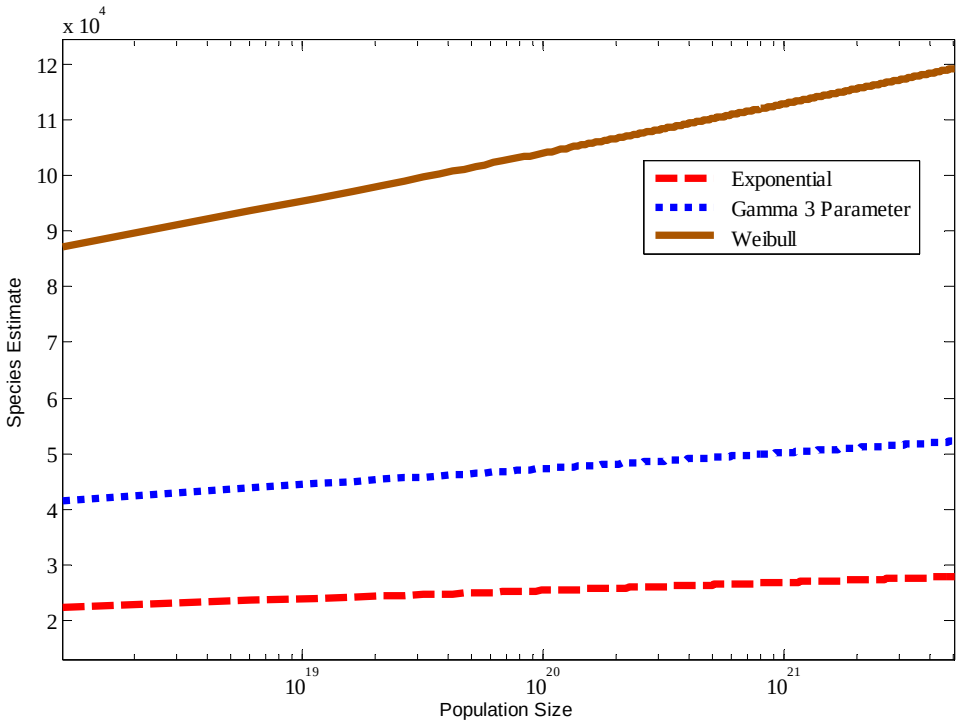

Figure 42. Parametric Estimates for Various Wetlands Population Sizes (Log Scale)

Fitting of three parametric curves to population sizes ranging from 1 x 10^{15} to 5 x 10^{21}. From Figure 41 it appears that each of the curves becomes asymptotic and, for any local area, that is the case. However, Figure 42 shows the same graph replacing the linear scale with a log scale; it is apparent that the curves do not truly become asymptotic. Still, although non-trivial assumptions must be made regarding the underlying population of the wetlands soil, in order to obtain similar results using the non-parametric methods, sample sizes much larger than the 2,820 provided for this research must be obtained. Therefore, the use of parametric methods for estimation should be considered as a better alternative to the non-parametric methods.

Research Questions Answered

The research question as taken from Chapter 1 is: Are the non-parametric methods currently used for species estimation, in fact, appropriate for use on microbial wetlands soil population estimation or does a better alternative exist?

This first part of this question was answered in the process of this research. In fact, the commonly used methods do not appear to be appropriate for application to this type of problem. This research also began to answer the second part of the research question. The use of parametric-curve fitting methodologies appears to hold more promise in attempting to accurately estimate the number of species comprising a large population.

Summary

This chapter delves into the analysis and results of the different data sets used in this thesis. An in-depth look at several contrived populations analyzed both non-parametrically and parametrically is offered. This is followed by the application of the methodologies to a real-world data set. This chapter concludes by presenting an answer for the first part of the research question, the commonly used non-parametric methods may not always be appropriate to use as estimators in this type of problem.

V. Conclusions and Recommendations

Chapter Overview

The purpose of this chapter is to provide the conclusions and recommendations brought about by this research. The overarching question to be answered concerns the appropriateness of applying non-parametric methods to large microbial populations for the purpose of species estimation. Furthermore, this thesis begins the process of identifying alternative methods to non-parametric estimation with regard to population composition. Conclusions of this research will be followed by its significance. This chapter will close with considerations for future research.

Conclusions of Research

Prior to the onset of this research, the question was posed as to whether the methods currently being used for species estimation were underestimating the true number of species in a very diverse population. The populations in question are very large, with the current thought suggesting at least 3,000,000 bacteria living in one gram of soil (Martin and Foch; 1977). To complicate the problem further, the samples which are acquired often are sparse (comprised of mostly singleton sequences). There is currently so little known about the truth that this is an extremely difficult problem. This research attempted to overcome the difficulty by creating several populations in which the truth regarding the number of species was known and applying the commonly used methods to samples from those populations. The results from those particular populations showed that the non-parametric estimators consistently failed to obtain the

true number of species. In every instance, the Chao1 method, the only non-parametric method for which the variance can easily be found, produced results which were many standard deviations from the truth. Therefore, caution should be exercised when applying this method as an estimator. Chao stated in her introduction to this method that it provides a lower bound for the number of species in the population, but shows promise as use for an estimator (Chao; 1987). This research questions the use of the lower bound as an estimator for the number of species. It is important to note that the results which are produced by continued use of the Chao1 method are producing lower bounds, not estimates. Further, these lower bounds are dependent on the sample size drawn from the population in question, as shown in Chapter 4 of this thesis. Additionally, even with an increase in sample size, Figures 10 and 11 show that the species estimate will either become essentially asymptotic. This is indicative that when a certain sample size is attained, a further increase in sample size will not result in a drastic improvement of a species estimate obtained by using the non-parametric methods discussed in this thesis.

In most instances, parametric fitting of curves to the data produced results which were closer to the truth. Although there are assumptions which must be made about the underlying population when using a parametric estimator, the results of parametric estimation show more promise than the use of non-parametric methods which only produce lower bounds. They consistently produced a lower percent error with regard to the actual number of species found in a contrived population than the non-parametric methods. However, the assumptions made with the use of a parametric method are not trivial assumptions. The two major assumptions that must be made are that of distribution type and that the population size is known. However, Chapter 4 of this thesis

demonstrated that population size will become a minimal assumption as the population size increases. Figure 14 showed that as the population size increased by orders of magnitude, the species estimate became virtually asymptotic. Since the estimated population size of the wetlands data is between 3.58×10^{15} and 5.97×10^{18} it may be that the more important assumption to focus on would be that of underlying distribution type. This hypothesis is further strengthened by noticing that although the population size increased by three orders of magnitude, there was only, on average, a 22 percent increase between the lower and upper confidence intervals for the parametric species estimates.

Significance of Research

This research provides a significant result for the biologists seeking to answer questions regarding species composition. The current thought is that the non-parametric estimators underestimate the number of species in a population. In reality, the non-parametric methods should not be used as estimators, but rather as lower bounds. Researchers wishing to continue using such methods should note that these methods are not acceptable for use as estimators and caution should be used when reporting results obtained from these methods.

Furthermore, this research indicated that even with an order of magnitude increase in sample size, the non-parametric species estimate did not experience the same increase. Essentially, there exists a sample size such that the estimated number of species in the microbial population will become essentially asymptotic. As such, a further increase in the sample size will not result in a much improved non-parametric species estimate.

Questions surrounding species composition continue to ply the scientific world. The area of species estimation continues to be prominent in scientific papers and journal articles. New methods for estimation are regularly being evaluated for use. It appears from this research that parametric methods hold more promise as estimators than do non-parametric methods.

Recommendations for Further Research

While this research answered the question surrounding the appropriateness of using the most common non-parametric methods, there are still questions which need to be addressed.

More research needs to be done into the area of parametric fitting of the data sets. Time should be devoted to further refining thoughts regarding the underlying distributions that nature follows. It may be that there is currently not a distribution which accurately describes the microbial population. In other words, perhaps the population distribution is unlike any which are currently being used (or considered for use) to describe the microbial world.

Additionally, it may be that no single distribution explains the microbial world and combinations of known distributions (mixed distributions) would be most appropriate to describing such populations. Attempting to mix different distributions may produce unexpected and/or unusual results.

Another interesting research area is investigating the possibility of making the parametric methods more invariant to population assumptions. Advancements in this

area would further strengthen the argument for using parametric methods over non-parametric methods.

A more in-depth study of the sample size and number of unique species and singletons in a given sample also merits further consideration.

Summary

Included in this chapter were conclusions and recommendations. The answer to the first part of the research question was presented. Progress was made toward answering the second. This research demonstrates that non-parametric methods should be considered as lower bounds (as originally intended) for population composition, not as estimators. Further, this research showed that parametrically fitting well-known probability distributions to the data sets provided answers which were closer to the truth. This chapter also offers suggestions for future research.

Appendix A: Matlab Code for Creation and Categorization of Contrived Populations

```
%% First define what you want the gamma function to be.  The parameters are
called A (alpha) and B(beta).  These parameters can be determined using
the disttool and adjusting the parameters.  For my first attempt I will
use A=0.8989 and B=1.095 (distribution can be seen using disttool).
loop through the following function.  I used each i as the nmber species (e.g. i=1
is the first species, i=2 the second, etc.)
format long
intmax('uint64')
clear C
clear D
A = input ('enter the value for the alpha parameter')
B = input ('enter the value for the beta parameter')
% The gamma function is a built in Matlab function.
Y=gamma(A)
X=((1/(Y*(B^A))))
X=X*10^9
%remember to change the ending index number to correspond to a small enough
density
for i=1:13484

   C(i)=(1/(Y*(B^A))*(i^(A-1))*exp(-i/B));
end
C=C';
%this gives the basic categories for the population (how many species in
population)
C=C*10^9
%the following gets the total population size
P=sum (C);
P=roundn(P,0)
%This is how to get number of species in each category
D=cumsum(C);
D=roundn(D,0);

%%
%this will sort the random sample into the corresponding species.  Remember
%to change the index to reflect the size of the random sample
for i=1:2840
   S(i)=sum(D<=W(i));
end
S=S'
sort(S)
```

Appendix B: Matlab Code for Non-Parametric Estimators

```
%%Chao estimator
% S is equal to the number of distinct species caught in the sample.
S = input ('enter the distinct number of species caught in sample')
%f1 is equal to the number of singletons found in the sample
f1 = input ('enter the number of singleton species found in sample')
%f2 is equal to the number of doubletons found in the sample
f2 = input ('enter the number of doubleton species found in the sample')
%Nchao is the estimation for the number of species based upon the above inputs
Nchao=S+(f1^2)/(2*f2)
%%estimation of the variance
Vchao=f2*[.25*((f1/f2)^4)+((f1/f2)^3)+.5*((f1/f2)^2)]
%% Aces estimator
% S is equal to Srare + Sabund
%Srare is the number of species with less than 10 members
Srare = input ('enter the number of species containing less than 10 members')
%Sabund is the number of species with greater than 10 members
Sabund = input ('enter the number of species containing more than 10 members')
n1 = input ('enter the number of species containing one individual')
n1 = 1*n1
n2 = input ('enter the number of species containing two individuals')
n2 = 2*n2
n3 = input ('enter the number of species containing three individuals')
n3 = 3*n3
n4 = input ('enter the number of species containing four individuals')
n4 = 4*n4
n5 = input ('enter the number of species containing five individuals')
n5 = 5*n5
n6 = input ('enter the number of species containing six individuals')
n6 = 6*n6
n7 = input ('enter the number of species containing seven individuals')
n7 = 7*n7
n8 = input ('enter the number of species containing eight individuals')
n8 = 8*n8
n9 = input ('enter the number of species containing nine individuals')
n9 = 9*n9
n10 = input ('enter the number of species containing ten individuals')
n10 = 10*n10
% Nrare is the total number of individuals within rare species.
Nrare = n1 + n2 + n3 + n4 + n5 + n6 + n7 + n8 + n9 + n10
%Ch is the abundance coverage estimator (ACE)
Ch= 1-(f1/Nrare)
% gamma is the coefficient of variation used in the calculation
```

```
gammace = (Srare/Ch)*((((1-0)*n1)+((2-1)*n2)+((3-1)*n3+((4-1)*n4)+((5-1)*n5)+((6-
1)*n6)+((7-1)*n7)+((8-1)*n8)+((9-1)*n9)+((10-1)*n10))/(Nrare*Nrare-1))-1)
if (gammace > 0) gammace = 0
end
Saces = Sabund + (Srare/Ch) + ((f1/Ch)*(gammace^2))
%%Jacknife estimator (Sjack1 is the first-order, focusing on singeltons in sample and
Sjack2 is the secong order, focusing on singletons and doubletons)
M = input ('enter the total sample size')
Sjack1 = S + (f1*((M-1)/M))
Sjack2 = S +(((f1*(2*M-3))/M)-((f2*(M-2)^2)/(M*(M-1))))
```

Appendix C: Matlab Code for Parametric Methods and Discretizing of Distributions

```
%%gamma 3-parameter

Alpha=input('enter value for alpha')
Beta=input ('enter value for beta')
Gamma=input ('enter value for gamma')
G=gamma(Alpha)
%pdf = (((x-Gamma)^(alpha-1))/((Beta^alpha)*G))*exp(-(x-Gamma)/Beta)
for i =1:50000
   H(i)=(((i-Gamma)^(Alpha-1))/((Beta^Alpha)*G))*exp(-(i-Gamma)/Beta);
end
H=H'
Hex=H*3850000000000000
Hex=Hex'
%%
%%discretize

a=1;
b=input('enter the endpoint for the summation')
n=input('enter the number of desired subintervals, should be equal to the number of
unqiue species found in sample')
h=(b-a)/n;
for i=1:b
   Hd(i)=((((a+i*h)-Gamma)^(Alpha-1))/((Beta^Alpha)*G))*exp(-((a+i*h)-
Gamma)/Beta);
end
Hd=Hd'
%%
a=1;
b=input('enter the endpoint for the summation')
n=input('enter the number of desired subintervals, should be equal to the number of
unqiue species found in sample')
h=(b-a)/n;
for i=1:b
   zd=((a+i*h)-Zt)/Lm
   Jd(i)=(Dt/((Lm*sqrt(2*pi)*z*(1-z))))*exp(-.5*(Gm+Dt*log(z/(1-z)))^2);
end
  Jd=Jd'

%%
%Lognormal
%sigma is a continuous parameter, greater than zero
%mu is a continuous parameter
%gamma is a continuous location parameter (is zero for two-parameter
```

```
%lognormal)

Sig=input('enter value for sigma parameter')
Mu=input('enter vaule for mu parameter')
Ga=input ('enter value for for lognormal gamma parameter, put zero if two parameter')
for i=1:250000

   K(i)=(exp(-.5*((log(i-Ga)-Mu)/Sig)^2))/((i-Ga)*Sig*sqrt(2*pi));
end
K=K'
Kex=K*3850000000000000
Kex=Kex'
%%
%%discretize
a=0;
b=input('enter the endpoint for the summation')
n=input('enter the number of desired subintervals, should be equal to the number of
unqiue species found in sample')
h=(b-a)/n;
%%
%%discretize
for i=1:b
    Kd(i)=(exp(-.5*((log((a+i*h)-Ga)-Mu)/Sig)^2))/(((a+i*h)-Ga)*Sig*sqrt(2*pi));
end
Kd=Kd'
RSlognormal=h*sum(Kd)
%%  Weibull
% alpha is a continuous shape parameter, greater than 0
% beta is a continuous shape parameter, greater than 0
% gamma is a continuous location parameter, equal to 0 for the two
% parameter distribution
alpha=input('enter the value for the Weibull alpha parameter')
beta=input('enter the value for the Weibull beta parameter')
gamma=input('enter the value for the Weibull gamma parameter, enter 0 if two parameter
distribution')
for i=1:70000
   L(i)=(alpha/beta)*(((i-gamma)/beta)^(alpha-1))*exp(-((i-gamma)/beta)^alpha);
end
L=L'
Lex=L*3850000000000000;
Lex=Lex'


%% Exponential
%lambda is the continuous inverse scale parameter, greater than 0
```

```
%gamma is the continuous location paramter, equal to zero for the one
%parameter distribution
lambda=input('enter the value for the exponential lambda parameter')
gammaE=input('enter the value for the exponential lambda parameter, enter 0 if one
parameter distribution')
for i=1:50000
   M(i)=lambda*exp(-lambda*(i-gammaE));
end
M=M'
Mex=M*3850000000000000;
Mex=Mex'
```

Appendix D: Probability Distribution Functions and Parameter Estimates

Exponential

$$f(x) = \lambda e^{(-\lambda(x-\gamma))}$$

where $\lambda > 0$ and λ is a continuous inverse scale parameter and γ is a continuous location parameter ($\gamma \equiv 0$ for the one-parameter probability distribution function).

Gamma

$$f(x) = \frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} e^{(-(x-\gamma)/\beta)}$$

where $\alpha > 0$ and α is a continuous shape parameter, $\beta > 0$ and β is a continuous scale parameter, and γ is a continuous location parameter ($\gamma \equiv 0$ for the two-parameter probability distribution function).

Lognormal

$$f(x) = \frac{e^{\left(-\frac{1}{2}\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^2\right)}}{(x-\gamma)\sigma\sqrt{2\pi}}$$

where μ is a continuous parameter, $\sigma > 0$ and σ is a continuous parameter, and γ is a continuous location parameter ($\gamma \equiv 0$ for the two-parameter probability distribution function).

Weibull

$$f(x)=\frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}e^{\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)}$$

where $\alpha>0$ and α is a continuous shape parameter, $\beta>0$ and β is a continuous scale parameter, and γ is a continuous location parameter ($\gamma\equiv 0$ for the two-parameter probability distribution function).

Parameter Estimates for Population 1

Rank	**Distribution**	**Parameters**
4	Exponential	λ=0.00696
3	Exponential (2P)	λ=0.007 γ=1.0
5	Gamma (3P)	α=0.89871 β=157.27 γ=1
1	Weibull	α=0.98183 β=145.09
2	Weibull (3P)	α=0.95576 β=143.84 γ=1

Parameter Estimates for Population 2

Rank	**Distribution**	**Parameters**
1	Exponential	λ=0.08905
5	Gamma	α=1.1488 β=9.7753
4	Gamma (3P)	α=0.47391 β=18.577 γ=1.0
2	Lognormal	σ=1.1922 μ=1.8407
3	Weibull	α=0.82309 β=10.055

Parameter Estimates for Population 3

Rank	Distribution	Parameters
5	Gamma	α=0.37297 β=139.27
1	Gamma (3P)	α=0.34649 β=102.34 γ=1.0
2	Lognormal	σ=1.8171 μ=2.5552
4	Weibull	α=0.53471 β=30.253
3	Weibull (3P)	α=0.38564 β=25.605 γ=1.0

Parameter Estimates for Population 4

Rank	Distribution	Parameters
5	Gamma	α=0.48268 β=369.89
1	Gamma (3P)	α=0.2928 β=613.04 γ=1.0
2	Lognormal	σ=2.2231 μ=3.5195
3	Weibull	α=0.50054 β=106.87
4	Weibull (3P)	α=0.3775 β=42.132 γ=1.0

Parameter Estimates for Wetlands Soil Data

Rank	Distribution	Parameters
3	Exponential	λ=0.00158
5	Exponential (2P)	λ=0.00158 γ=1.0
1	Gamma (3P)	α=0.5236 β=1267.5 γ=1.0
4	Lognormal (3P)	σ=1.6338 μ=5.6328 γ=-6.8803
2	Weibull	α=0.70809 β=599.87

Bibliography

Bishop, Ethan. *Molecular Characterization of Wetland Soil Bacterial Community in Constructed Mesocoms*. MS thesis, AFIT/GES/ENV/06J-01. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, June 2006 (ADA466099).

Burnham, K.P. and Overton, W.S. "Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals," *Biometrika*: 625-633 (1978).

Chao, Anne. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability," *Biometrics*: 783-791 (1987).

Chao, Anne. "Estimating Population Size for Sparse Data in Capture-Recapture Experiments," *Biometrics*: 427-438 (1989).

Chao, Anne and Lee, Shen-Ming. "Estimating the Number of Classes via Sample Coverage," *Journal of the American Statistical Association*: 210-217 (1992).

Chao, Anne and Yang, Mark. "Stopping Rules and Estimation for Recapture Debugging with Unequal Failure Rates," *Biometrika*: 193-201 (1993).

Chao, Anne and Lee, Shen-Ming. "Estimating Population Size via Sample Coverage for Closed Capture-Recapture Models," *Biometrics*: 88-97 (1994).

Chao, Anne and Shen, Tsung-Jen. "Nonparametric Estimation of Shannon's Index of Diversity when there are Unseen Species in Sample," *Environmental and Ecological Statistics*: 429-443 (2003).

Chao, Anne. "Species Richness Estimation," Paper. National Tsing Hau Univeristy: 1-23 (2006).

Collins et al. eds. *The Significance and Regulation of Soil Biodiversity*. Netherlands: Kluwer Academic Publishers, 1995.

Colwell, Robert and Coddington, Jonathan A. "Estimating Terrestrial Biodiversity Through Extrapolation," *Phil. Trans. R. Soc. Lond*: 101-118 (1994).

Dove, Alistair and Cribb, Thomas. "Species Accumulation Curves and their Applications in Parasite Ecology," *Trends in Parasitology*: 568-574 (2006).

Hong et al.. "Predicting Microbial Species Richness," *PNAS*: 117-122 (2006).

Hughes et al.. "Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity," *Applied Environmental Microbiology*: 4399-4406 (2001).

Kemp, Paul and Aller. Josephine. "Bacterial Diversity in aquatic and Other Environments: What 16s rDNA Libraries can Tell Us," *FEMS Microbiology Ecology*: 161-177 (2003).

Kemp, Paul and Aller, Josephine. "Estimating Prokaryotic Diversity: When are 16s rDNA Libraries Large Enough," *Limnology and Oceanography*: 114-125 (2004).

Leon, Elisabeth. *Molecular Characterization of Wetland Soil Bacterial Communities in Constructed Mesocosms*. MS thesis, AFIT/GES/ENV/08-M04. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2008.

National Institute of Standards and Technology. "Kolmogorov-Smirnov Goodness-of-Fit Test," *Engineering Statistics Handbook*. 25 February 2008 http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

Martin and Focht. in Elliott and Stevenson, eds. "Soils for Management of Organic Wastes and Waste Waters," *American Society of Agronomy*. (1977).

Offwell Woodland and Wildlife Trust. *Simpson's diversity Index* (August, 2007). http://www.countrysideinfo.co.uk/

Vita

Captain Melanie R. Slattery graduated from Middletown High School North in Middletown, NJ. She graduated with an undergraduate degree in Mathematics, Bachelor of the Arts, from Rowan University in Glassboro, New Jersey in May 1997. She was accepted as a candidate for Officer Training School in August 2003 and received her commission in October 2003.

Her first assignment was at Tyndall, AFB with the 83rd Fighter Weapon Squadron. There she served as the squadron executive officer and as an AIM-9 missile analyst. In August 2006, she entered the Graduate School of Engineering and Management, Air Force Institute of Technology. Upon graduation, she will be assigned to the United States Air Force Academy Preparatory School.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
03-27-2008	**Master's Thesis**	Jun 2008 – Mar 2008

4. TITLE AND SUBTITLE

Estimation of the Number of Microbial Species Comprising a Population

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Slattery, Melanie, R., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GAM/ENC/08-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally left blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The purpose of this research was to evaluate the appropriateness of using non-parametric estimators, specifically the Chao1, ACES, and Jackknife methods, for estimation of the number of unique species comprising a population. This research consisted of creating diverse populations, with a known number of species, and applying the aforementioned methods to samples drawn from the constructed populations. The analysis of the non-parametric methods was followed by the parametric fitting of several different distributions to the sample data, including the lognormal, gamma, and Weibull. These results were analyzed as well. Both types of methodologies were then applied to sample data from constructed wetlands, where little is known about the population size and composition. This research did not attempt to identify the underlying population distribution of the wetlands, but rather focused upon demonstrating that the use of parametric methods are more apt to provide better results in estimating the number of species in a natural population.

This research discovered the use of the non-parametric methods is not an appropriate for species estimation. The use of these methods resulted in lower bounds, which were several standard deviations away from the true number of species, for the contrived populations. A parametric method was more accurate in representing the truth. Recommendations for further research are provided in this thesis.

15. SUBJECT TERMS

Parametric Estimation of Species, Chao1 Method, ACES Estimator, Non-parametric Estimation of Species, Prokaryotic 16s rDNA Libraries

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
REPORT	**ABSTRACT**	**c. THIS PAGE**			Samuel A. Wright, Maj, USAF (ENC)
U	**U**	**U**	**UU**	108	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4549; e-mail: Samuel.Wright@afit.edu

Standard Form 298 (Rev: 8-98)
Prescribed by ANSI Std. Z39-18